A semantic framework for social search

Johann Stan

UNIVERSITY OF LYON - UNIVERSITÉ JEAN-MONNET
DOCTORAL SCHOOL Sciences, Ingénierie, Santé
(ED 488)

P H D T H E S I S

to obtain the title of

Doctor in Computer Science

of the University Jean-Monnet
Specialty : Computer Science

Submitted by
Johann STAN

A Semantic Framework for Social Search

A Semantic Framework for Social Search

In recent years, online collaborative environments, e.g. social content sites (such as Twitter or Facebook) have significantly changed the way people share information and interact with peers. These platforms have become the primary common environment for people to communicate about their activity and their information needs and to maintain and create social ties. Status updates or microposts emerged as a convenient way for people to share content frequently without a long investment of time. Some social platforms even limit the length of a "post". A post generally consists of a single sentence (e.g. news, a question), it can include a picture, a hyperlink, tags or other descriptive data (metadata). Contrarily to traditional documents, posts are informal (with no controlled vocabulary) and don't have a well established structure.

Social platforms can become so popular (huge number of users and posts), that it becomes difficult to find relevant information in the flow of notifications. Therefore, organizing this huge quantity of social information is one of the major challenges of such collaborative environments. Traditional information retrieval techniques are not well suited for querying such corpus, because of the short size of the share content, the uncontrolled vocabulary used by authors and because these techniques don't take in consideration the ties in-between people. Also, such techniques tend to find the documents that best match a query, which may not be sufficient in the context of social platform where the creation of new connections in the platform has a motivating impact and where the platform tries to keep on-going participation. A new information retrieval paradigm, social search has been introduced as a potential solution to this problem. This solution consists of different strategies to leverage user generated content for information seeking, such as the recommendation of people. However, existing strategies have limitations in the user profile construction process and in the routing of queries to the right people identified as experts. More concretely, the majority of user profiles in such systems are keyword-based, which is not suited for the small size and the informal aspect of the posts. Secondly, expertise is measured only based on statistical scoring mechanisms, which do not take into account the fact that people on social platforms will not precisely consume the results of the query, but will aim to engage into a conversation with the expert. Also a particular focus needs to be done on privacy management, where still traditional methods initially designed for databases are used without taking into account the social ties between people.

In this thesis we propose and evaluate an original framework for the organization and retrieval of information in social platforms. Instead of retrieving content that best matches a user query, we retrieve people who have expertise and are most motivated to engage in conversations on its topics. We propose to build dynamically profiles for users based on their interactions in the social platform. The construction of such profiles requires the capture of interactions (microposts), their analysis and the extraction and

understanding of their topics. In order to build a more meaningful profile, we leverage Semantic Web Technologies and more specifically, Linked Data, for the transformation of microposts' topics into semantic concepts. Also, we introduce an original profile scoring mechanism for the quantification of expertise.

In particular we investigate how (i) to transform social content into semantic concepts that have an ontology-based representation; (ii) to design a social search framework that takes full advantage of the rich semantics of such representations; (iii) to capture the expertise of the user according to the style and the content of the messages; (iiii) manage the privacy of user profiles, allowing defining granular variations of concepts to be shared with a particular social category. We propose then a toolkit for the semantic exploration of online communities and a new user model based on the content productions of users.

Our thesis contributes to several fields related to the organization, management and retrieval of information in collaborative environments and to the fields of social computing and human-computer interaction.

Context of the Thesis: Alcatel-Lucent Bell Labs France - Social Communications Department

This thesis was performed in the Social Communications Department of the Applications Domain of Alcatel-Lucent Bell Labs France [1].

The Applications Domain is composed of several departments that produce scientific and technological assets that will be integrated in the next generation products of Alcatel-Lucent. One of these departments, Social Communications, studies different ways of easing the communications between people in their specific social context.

More specifically, the objective of this department is to propose a middleware between the physical and the digital world in order to offer users the best of both worlds in any possible social context. In order to achieve this, several research challenges are addressed:

- the collection, analysis and exploration of social data shared by users in the social ecosystem (e.g. social networks, social bookmarking systems, content sharing systems etc.) and the design of advanced recommendation strategies for helping users in decision making
- the assistance of users in real time in their social activities (e.g. during the lecture of a book).
- the design of new objects that can help users better communicate and better find the right information for a specific problem

Performing this thesis in an industrial context allowed me not only to publish scientific publications about my work, but also to participate in other projects and submit patent applications. Therefore, most of the components of this framework are also submitted as patent applications. Also, performing the thesis in this context allowed me to have the necessary motivation to develop a fully functional social search engine and to map my work on the current research strategy in Bell Labs.

[1] `http://www.alcatel-lucent.com/wps/portal/BellLabs/`

Acknowledgments

This thesis was performed in the frame of a CIFRE[2] convention between Alcatel-Lucent Bell Labs France and Laboratoire Hubert Curien, Jean-Monnet University.

When I decided to start a PhD thesis, I knew it would be difficult, challenging, but also a very enriching personal and professional experience. As I am now looking back to these years, I realize how important it was to be supported by my suporvisors, colleagues, friends and family. This section is dedicated to them.

First of all, I would like to express my gratitude to Dr. Pierre Maret, my research supervisor, for giving me the opportunity to work with him and his wise guidance, advice and suggestions. The different visits we made together to research institutes in Germany allowed me to present my work in numerous occasions and to receive valuable feedback from experts in the fields.

I would like to give very special thanks to Elod Egyed-Zsigmond, who was my first supervisor in research in LIRIS, Lyon. By working with him I learnt the fundamental skills that are necessary for top-level research, from addressing a scientific problem to writing a paper.

My most sincere thank you to my former colleague in Bell Labs, Dr. Adrien Joly, who hired me as an intern to perform my Master thesis in the Social Communications department of Bell Labs. The Head of the department, Johann Daigremont, helped me a lot to progress my research talks and presentations and I would like to express my gratitude for this help. It was him who gave me the necessary resources and autonomy to carry out my thesis in the best possible conditions.

I acknowledge also the help provided by Fabien Bataille and Claire Le Foch who acted as patent reviewers during my work in Bell Labs. His precise reviews allowed me to gain significant experience in writing patents and defending them.

Special thanks to my former colleagues, Lionel Natarianni, Julien Robinson, Mathieu Beauvais and Sylvain Squedin who always trusted and supported my work. I am also grateful to all members of Bell Labs for the interesting discussions I had on various research topics.

I would like to specially thank Myriam Ribière, for receiving me in her team during a difficult time and helping me finding a new path in my thesis.

It was also a great pleasure to supervise Viet-Hung Do, who successfully performed his Master Research internship under my guidance in Bell Labs.

Last but not least I thank my family for their help, support and love in these years. This thesis is dedicated to them.

[2]Conventions Industrielles de Formation pour la REcherche

Contents

Part I

Introduction and State or the Art

Chapter 1

Introduction

Contents

1.1 Introduction and Objectives

In recent years, online collaborative environments, e.g. social content sites, also called social platforms [Amer-Yahia 2009a] (i.e. systems that encourage users to share social information and engage in interactions, e.g. Twitter[1] and Facebook[2]) have significantly changed the way people organize, share information and interact with peers. These platforms have become the primary common environment for people to communicate about their activity and their information needs, to maintain and create social ties. So called status updates or microposts emerged as a convenient way to share content frequently without a long investment of time. Some social content sites even limit the

[1] Twitter Microblogging Site - `www.twitter.com` - visited August 2011
[2] Facebook Social Network Site - `www.facebook.com` - visited August 2011

length of a "post". A post generally consists of a single sentence (e.g. news, a question), it can include a picture, a hyperlink, tags or other descriptive data (metadata). Contrarily to traditional documents, posts are informal (with no controlled vocabulary) and don't have a well established structure.

Mainly because of the simplicity of use, social content sites can become so popular (huge number of users and posts), that it becomes then difficult to find relevant information in the flow of activity notifications. Therefore, organizing this huge quantity of social information is one of the major challenges of such collaborative environments. Traditional information retrieval techniques are not well suited for querying such corpus, because of the short size of the shared content, the uncontrolled vocabulary used by authors and because these techniques don't take in consideration the ties in-between people. In other words, these techniques are not tailored to systems that integrate both content and social information.

Finding the documents that best match a query may not be sufficient in the case of a social platform where the creation of new connections between members is an important objective and where the platform should try to keep on-going participation of its members [Lee 2003] in order to grow. A new information retrieval paradigm, social search [Bao 2007] [Morris 2010] [Horowitz 2010] has been introduced as a potential solution to this problem. This solution consists of different strategies to leverage user generated content for information seeking, such as the retrieval of people who may have the right knowledge to answer a query. Indeed, recommending people to send a post to appears as more important as finding the right information in the social platform, because this information may not exist and because part of the added value relies into the conversation engaged in-between peers.

However, existing social search strategies have several limitations with regards to content management and information discovery. In order to implement such a search strategy, information is required about the interests and expertise of users, which must be captured from their activities and behavior in the platform, such as by analyzing the content they share. The particular nature of content shared in such systems gives birth to new challenges with regards to their understanding. In current approaches, the user profile is composed of keywords that are extracted from the messages. Given the fact that messages are short, this results in profiles that contain few information about the user. Finally, the weighting schema of the profiles is generally based on statistical scores, such as term frequency. The style of the shared content (e.g. vocabulary used to share about a given topic) is not considered to acquire additional knowledge on the author competences and intentions. To the best of our knowledge, there is no general approach that allows to implement such a search strategy on popular social content sites, allowing to benefit from trusted social resources, such as the knowledge of friends in case of an information need.

In this thesis we propose and evaluate a framework for the organization, retrieval

and exchange of information in social content sites. Instead of retrieving content that best matches a user query, we retrieve people who have expertise and are most motivated to engage in conversations on these topics. We propose to build dynamic expertise profiles for users based on their interactions in the social platform. The construction of such profiles requires the capture of interactions (microposts), their analysis and the extraction and understanding of their topics. In order to build a more meaningful profile, we leverage Semantic Web Technologies and more specifically, Linked Open Data [Lehmann 2009] knowledge bases, for the transformation of microposts' topics into semantic concepts. Also, we introduce a profile weighting mechanism for the quantification of expertise and interactivity, based on the analysis of statistical sharing patterns and the style of shared content.

In particular we investigate how (i) to transform social content into semantic concepts that have an ontology-based representation; (ii) to design a social search framework that takes full advantage of the rich semantics of such representations; (iii) to capture the expertise of the user according to the style and the content of the messages; (iiii) manage the privacy of user profiles, allowing defining granular variations of concepts to be shared within a particular social category. We propose and evaluate then a toolkit for the semantic exploration of online communities and a new user model based on the content productions of members of social platforms.

The main goal of this thesis is to contribute to information management and discovery in social content sites with a semantic framework allowing to implement a social search strategy that allows to discover interesting people for a question and to contribute to several fields related to the organization, management and retrieval of information in collaborative environments, as well as to the fields of social computing and human-computer interaction.

1.2 Scientific Challenges and Overview of our Approach

After the introduction of our goal in the previous section, in the following we inform on the main scientific challenges that need to be considered in order to build the semantic social search framework with the objective to improve the organization and discovery of knowledge in social content sites.

1.2.1 From Microposts to Semantic Concepts

As we mentioned before, in order to capture user's expertise, we must follow his/her activity and behavior. In the case of a social content site, the most frequent activity is the sharing of content, which can be captured and analyzed. The main novelty in our approach for the construction of expertise profiles is the use of content productions instead of content consumptions.

However, the particular style of content productions in such platforms results in the fact that generally few useful information can be extracted (e.g. keywords, named entities). Our solution to enrich the quantity of information we have about a user's interests is to leverage semantic knowledge bases in the frame of Linked Open Data. The size limitation imposed by the majority of these platforms requires users an additional effort to formulate their message as concisely as possible. This results in the fact that they will share only a fragment or summary of their thought. The connection of the topics of these messages to a semantic knowledge base allows to enrich it with additional concepts and thus, enlarge the digital representation of the user's interests.

The main challenge in this case is to find the concept in the knowledge base that best matches the meaning of the message. Particularly in our case, we lack contextual cues to efficiently disambiguate these topics, which can be either keywords or named entities in the message. Therefore, in order to enrich the available contextual cues, we take into account two additional elements: (i) the previous messages of the user and (ii) the community of the user. Once the right concept is identified in the knowledge base, we perform an operation called semantic expansion in order to retrieve additional concepts that are potentially relevant for the user. For example, when a user expresses an opinion about a movie, the names of actors can be good candidates for the enrichment of the profile, as expertise and interactivity can be further propagated to these concepts. This operation allows to have richer profiles and correspondingly, increases the probability to find the best expert for a query. We leverage the structure of the knowledge base for the computation of similarities between the concepts of a query and user profiles and for an improved privacy management of the profile, by allowing users to share granular variations of a given concept with different social categories.

1.2.2 The Identification and Quantification of Expertise and Interactivity

Producing content requires an additional cognitive effort from users and it reflects better their interests and expertise. Also, in a content production we can mine for additional information regarding users' state of mind when performing the content sharing task, such as their sentiment. The vocabulary they employ when sharing information in a given domain is a good indication of the underlying expertise. Statistical measures, temporal patterns can bring further insights in their expertise and motivation to engage

in an interaction on a specific topic.

1.2.3 Privacy. Trust. Intimacy

Replacing documents with people in the search strategy gives birth to an additional difficulty, which is represented by trust and intimacy, inexistent in the case of traditional document-based information retrieval. It has been demonstrated that a user trusts more recommendations from friends than from more distant connections [Mendes 2010a]. However, there are situations where explicit friends do not know the answer to a question. It is thus important to explore more distant connections in the social platform, as this could result in new connections between people.

1.3 Summary and Classification of Contributions

This work contributes to several scientific domains in the area of computing: (i) Social Network Analysis, (ii) User Modeling and (iii) Privacy Management. In this section we briefly introduce these domains and the corresponding contributions.

1.3.1 Social Network Analysis

Social Network Analysis is a subdomain of Social Information Processing, part of the general domain of Social Computing. This is a general term for an area of computer science that is concerned with the intersection of social behavior and computational systems [Wang 2007]. Although the majority of contributions in social network analysis deal with the understanding of the topology of a social platform, more and more work is taking into account the analysis of the shared content in such a network, as this can significantly help in detecting meaningful communities or to extract the interests of a user [Wagner 2010].

We contribute to this field with a framework and algorithms that allow to perform the semantic analysis of messages shared in social content sites, which all have an underlying social network, and in particular, the matching of the topics of such messages to a semantic knowledge base. Such disambiguated concepts can then provide a valuable input for advanced recommendation strategies targeted to such systems [Stan 2011a].

1.3.2 User Modeling

User modeling is a sub-area of human - computer interaction (HCI), which defines cognitive models of human users, including modeling of their skills and declarative knowl-

edge. Numerous applications of such systems exist for example in the area of natural language understanding and dialogue systems, in computer-based educational systems and online learning environments, in systems for computer supported collaboration and recommender systems. A user profile is the digital representation of such a user model. Our objective with regards to this field is to provide a user model that is built by taking into account the content the user explicitely shares in a platform. In this way, the user model will reflect the cognitive effort of the user and will contain items that the user considered worth sharing about. In the case of a content consumption (e.g. visiting a web page), few information can be captured about the interest of the user. For example, one can capture the time the user spent on the web page or the different activity he/she performed while visiting it (e.g. scrolling with the mouse).

In the case of a content production, we can take a deeper insight into the state of the mind of the user: we can extract his/her sentiment or we can analyse the style of the message. Such additional information may be of great value for e.g. capturing the expertise of the user. This is the main reason we consider content productions a more interesting input for the construction of a user expertise profile in a social platform.

Therefore, our contribution to User Modeling is multifold: (i) the definition of a user profile model based on content productions of the user in a socila platform, (ii) the definition of a user model that is composed of concepts from Linked Open Data and (ii) the definition of a scoring function that allows to capture a rich description of the user's expertise for a given profile concept [**?**] [Robinson 2011].

1.3.3 Privacy Management

Privacy is a fundamental component of human-computer interaction and user modeling. A privacy management strategy defines how the user can regulate the diffusion of his/her private social information. [Barker 2009] defines a taxonomy for data privacy management and one of the dimensions is considered the granularity, which defines what a level of detail of a given data item will be shown to a given person, when he/she accesses it. However, to the best of our knoweldge, this dimension has not yet been translated for privacy management in social content sites.

Therefore, we contribute to this field with a new privacy management strategy that allows users to define granular variations of a profile concept for a particular social sphere. [Hacid 2010]

1.3.4 Social Search

Social Search refers to the retrieval of information using social resources. This search strategy can be further conceptualized with the library/village metaphor, mentioned by [Horowitz 2010]. As described by the authors, the traditional and basic paradigm for information retrieval has been the library, as Google itself has its origins in the Stanford Digital Library project [Page 1999].

This paradigm indeed worked well in the early ages of the Internet, where users had no possibility to contribute to the content of web pages, by commenting or giving feedback. However, this search strategy ignores the social dimension of content, which emerged from web 2.0 practices and communication processes, which puts the user in the center of information management.

This new dimension can be best illustrated with the metaphor of a village, where knowledge dissemination is achieved socially by word-of-mouth, meaning that information is passed from person to person and the retrieval task consists of finding not the right document, but the right person to answer an information need. In the case of social platforms, this second paradigm is more interesting for both the service provider and members of the platform.

There are two main reasons for this:

- a people-to-people search strategy may result in new connections in the system which is very important to its growth and attracktivity to newcomers
- identifying a user as an expert for a given information need may have a motivating impact to him/her for further content sharing, as the user may feel him/her-self rewarded by the system for the effort of sharing interesting content

We contribute to this field with an original framework [Stan 2011b] that can be integrated in a given social platform and that retrieves people most relevant to a user query expressed explicitly in natural language or implicitly, by a particular human-computer interaction, such as web navigation.

1.4 Motivation Scenarios

This section presents two motivational scenarios for our work. The first is a larger vision, related to the fact that currently the digital and physical lives of users are independent silos with few useful connections (i.e. rarely can users benefit from a discussion, conversation on a social content site in a real life situation). The second is more related to human organizations and to the reduction of the latency of knowledge exchange in such places.

1.4.1 Bridging the Gap between Physical and Digital Lives

Our first motivation to build a social search framework on top of existing social content sites is motivated by the vision of connecting the physical and digital lives of users. Being part of a social platforms allows users to be aware of the real-time activities of their connections from different social spheres and correspondingly, to have a rich community experience. However, the increase of shared content and correspondingly, activity streams, makes it more and more difficult to find useful information or to discover people with similar interests.

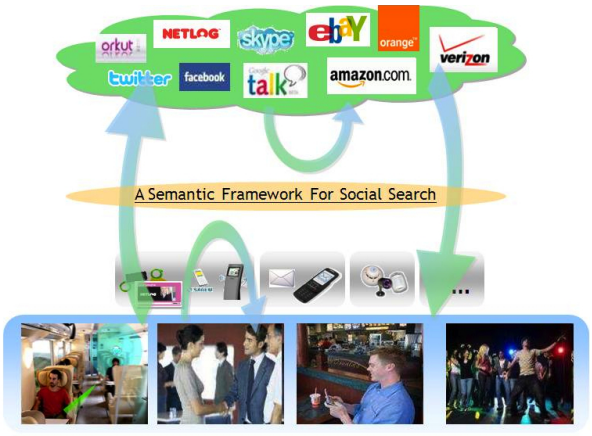
Our vision is therefore to provide a framework that connects the digital world of users (Figures 1.1(a) and 1.1(b)), composed of their life in social platforms with the real, physical world, composed of different situations where they might need assistance, such as traveling, being at a conference or searching a good doctor in a new city. Today, it is difficult to find people in a community who can be useful for such an information need. It would then require to search in the social updates of each connection hoping to find some useful information and this can be a very time-consuming task.

Such information needs can be expressed with a question in natural language.

Our vision can be also summarized as a capacity of a social platform to be semantically aware of members' interests and expertise and correspondingly, to be able to recommend a user the connections who are most relevant to their need of information. In this work, we provide a solution to this vision based on the previously mentioned user expertise identification in social content sites and a corresponding semantic framework for social search strategy.

1.4.2 Knowledge Latency in Human Organizations

In network theory, latency, synonym for delay, is an expression of how much time it takes for a packet of data to get from one designated point to another. Inspired by this concept, we introduce the vision of "Knowledge Latency", which translates the concept of latency from networks to human organizations. *More concretely, knowledge latency relates to the need for highly distributed organizations to share and spread more efficiently knowledge across borders.* Knowledge distinguishes from information by its human dimension: knowledge is information assimilated by human beings and therefore, knowledge latency in a human organization allows to measure the time needed for the transmission of knowledge from one person to another and its assimilation (i.e. capacity to have his/her own point of view on the received knowledge and to share it to other peers by using his/her proper vocabulary). An important challenge in such human organizations is to reduce the knowledge latency in order to gain in efficiency and reduce costs. A possible solution for its reduction could be the connection of people with different information needs to experts who can efficiently help them for the given problem.



(a) Connecting Digital and Physical Worlds

(b) Finding friends who can recommend a dentist in Paris

Figure 1.1: Our vision

The ideal situation in this case would be certainly that of "zero latency" defined in a white-paper by the Hewlett-Packard company[3]. In this case, this concept is discussed as a broader notion of all enterprise operations but has relevance to the terms and systems we are discussing and also addresses some rationalization processes that may need to be considered as part of achieving the experience we hope to provide the user. According to the authors of this white paper, in an enterprise context, Gartner was among the first in 1998 to put together the vision of the zero latency enterprise as the "instantaneous awareness and appropriate response to events across an entire enterprise." In the case of the underlying human organization in the enterprise, zero latency would certainly

[3]HP White Paper on Zero Latency - `http://h71028.www7.hp.com/ERC/downloads/ZLEARCWP.pdf`- visited September 2011

mean the seamless exchange of knowledge between knowledge workers and in our view, the first step in this process is having the capacity to rapidly find experts that are motivated to help in our information and communication needs.

This scenario looks passed the more traditional Query/Results user scenario and considers providing additional support by providing the user with pointers to human knowledge resources identified as such by possibly fuzzy associations. This helps us move beyond the simple (but not necessarily easy) Q&A session and helps build relationships between people to further reduce knowledge latency for the current query but also build relationships that may narrow these latencies for future queries. It may also stimulate the creation of new knowledge, as interacting with an expert or simply, new people, can result in new questions, new ideas, which is very important for the innovation process in human organizations (e.g. an enterprise or an educational institution etc.).

1.5 Dissertation Plan

In this section, we depict the sequence of research activities that have been undertaken from the problem statement introduced in this first chapter, towards the contributions discussed in chapter six. In the first part of the thesis (chapters two to four), we have carried out a state of the art in the different domains mentioned before that are relevant to address the theoretical and technical challenges towards the envisioned semantic framework for social search, and to position our contribution to existing work.

More specifically, the outline of this thesis is as follows.

First Part: State of the Art

- In Chapter 2, we introduce web-based social content sites, also called social platforms and define a common environment for them called "social web ecosystem". We focus on the specific characteristics and building blocks of these platforms. We analyze their usage through the enumeration of salient trends and major evolutions that we have observed since their appearance on the Internet. We define two main pillars for such systems, the users and social objects. It is thanks to the users that a social platform is alive, as they produce content and engage in diverse social interactions. It is therefore important to understand the dimensions of the digital life of a user. In addition, we define social objects as the second pillar, which is the content produced by the users. We identify in this case the different types of content shared in such platforms, and we investigate the reasons why social content have converged to short-sized microposts in recent years.

- In Chapter 3 we first introduce the scientific domains that are strongly connected to our challenge of building a social search framework for expertise identification and retrieval. Our work is directly connected to three major scientific domains: (i)

social computing - social network analysis, (ii) semantic web - semantic metadata management and (iii) human-computer interaction - user modeling and privacy management. In the case of social network analysis, we decompose this field in two main areas: structural social network analysis and semantic social network analysis. In the frame of semantic social network analysis, most research effort is done for the analysis of social tagging systems. However, recently, as social updates become more and more the most popular form of communication in such platforms, their analysis gains more attention. We also review in this chapter the literature in semantic metamodeling in order to depict the structure and scope of vocabularies that compose Linked Data. This review also helps us in understanding how social data is annotated with such structured vocabularies, which is necessary also for our framework. Finally, we review the two fields that are related to human-computer interaction: user modeling and privacy management. The objective of this first part of the state of the art is to define the perimeter of our work.

The last section of this chapter reviews existing social search frameworks and identifies work related to our proposal. We identify Aardvark as the closest system that has similar objectives [Horowitz 2010] and identify its limitations.

In the last part of this chapter, we discuss the state of the art and identify the main limitations of existing approaches that could compose our solution. We also identify existing algorithms that are necessary to build our framework and that can be easily adapted to this specific context.

Second Part: Contributions. Evaluation. Conclusion and Perspectives.

- In Chapter 4 we present our semantic framework for social search. The first section of this chapter presents an analysis of the structure and content of microposts, which allows to identify what extraction mechanism are required to extract their topics. Also, we identify the different user populations in such systems, from people who share very general information about their activity, to people who share more specific information about their findings and interests. The second section of this chapter presents the theoretical framework for social search and its main layers from the capture of social interactions, their analysis and the construction of user expertise profiles. The following section reports on each component. More specifically, the construction of a user profile is composed of two main steps:

  - The identification of user profile items (e.g. user interests) based on content productions. We present for this two algorithms for semantic matching (SoSeM) and semantic expansion (SemEx). These two algorithms allow to

manage the user profile on a conceptual level, by associating profile items to semantic concepts.

- The scoring mechanism for the association of weights to user profile items. Our scoring mechanism analyses the content shared by a user to extract a rich set of information that is useful for measuring the expertise, i.e. the sentiment and the complexity of used vocabulary.

The final contribution section presents a new approach for the management of the privacy of such user profiles. Indeed, the user of concepts from semantic knowledge bases allows to define a more flexible privacy policy, by considering different levels of detail of a given concept.

- In Chapter 5 we inform on our proof-of-concept prototype and describe our experimentation results and validate our approach with an experimentation protocol involving end-users.

To conclude this work, in Chapter 6, we discuss our contribution and the implied findings and limitations. Future work, recommendations and perspectives are then proposed in order to allow different professionals to continue research in this direction.

Chapter 2

The Anatomy of Social Platforms

Contents

In the early days of Web 2.0, also called "social web", the most important interaction and online communication tools were represented by online forums, blogs and content annotation platforms. Later on, platforms inviting users to engage in social interactions based on shared social content have become the most widely used communication paradigm on the Internet, enabling to define friend lists, user profiles and sharing content facilitating in this way the communication with others. From an upper-view, such a social platform can be viewed as a virtual community, in which users communicate using different virtual communication processes, such as exposing oneself with the help of a public profile, communicating one-to-one, one-to-many, many-to-many, sharing content, commenting any of the above and interaction between all of these.

In this chapter we present the general characteristics of social platforms and define their main pillars and respective dimensions. In other words, our objective is to define a general model for social platforms by identifying their most important building blocks.

2.1 The Anatomy of the Social Web Ecosystem

A social platform is generally forged by a virtual community of members. The concept of virtual communities (VCs) has its origins in the definition of systems composed of autonomous agents, capable of interactions for achieving collaboratively a goal [Camarinha-Matos 2004] [Maret 2004]. However, with the emergence of the first social

platforms, it was also used to define the underlying online community of social platforms [Wellman 1996].

In the rest of this chapter, we examine the origins and objectives of these communities and study their main pillars, providing a general model for them from the perspective of social platforms.

2.1.1 The Background of Virtual Communities (VC), as the Underlying Human Organization of Social Platforms

In this section, we first discuss the concept of virtual communities and its main pillars. Secondly, we examine the nature of content productions in these communities and corresponding challenges to manage and process social information.

2.1.1.1 Online Communities

Community is a term that continuously evolves through the change of technologies and human behavior. The concept of community exists also in sociology and is defined as a group of interacting people, living in a common location and organized around common values. Since the advent of the Internet, the concept of community has no longer geographical limitations. People can now virtually gather in online communities, forming virtual communities where they can share common interests regardless of physical location.

Since its beginning in the '70s, the Internet technologies in its early forms, like bulletin board systems (BBS), Internet Relay Chat (IRC), Usenet and forums have been strongly bound to the community concept. Historically, the Well ("The Whole Earth 'Lectronic Link")[1], a computer conferencing system that enables people around the world to carry on public conversations and exchange private electronic mail (e-mail) in 1985 appears as the first real online community.

At this early stage, online-communities were used to describe persons using Internet to make online discussions on a communication space that we used to call "the cyberspace".

The rapid development of online virtual communities really occurred after dot.com crashed in 2002, where the Internet and the Web was reinvented with the label Web 2.0. Since this period, user-generated content and social interactions in virtual spaces all over the world have characterized the main form of interactions. Social networks, such as Facebook or LinkedIn and more recently Twitter are commonly known as the most successful and visible part of online-communities. Today, with the development of mobility, real-time Web and pervasive computing online interactions become more and more transient and opportunistic.

[1] The Well System - `www.well.com` - visited August 2011

2.1.1.2 Participation in Virtual Communities

The term of virtual communities was used for the first time in 1993 by H. Rheingold in the book "The Virtual Community" [Rheingold 2000] and described people connected online, having long public discussions, which forged personal relationships. Providing tools for such communities have multiple interests for both businesses and individuals.

A growing number of companies are building VCs to facilitate peer-to-peer help [Constant 1994], foster new ideas and innovation [Nambisan 2002], and build knowledge competencies [Saint-Onge 2003]. Many firms are hosting online user communities to collect feedback and ideas [Williams 2000] and to strength, improve their innovation process [Jeppesen 2006].

In public sectors, VCs emerge to leverage the knowledge embedded in professionals, e.g., open-source communities and community of practices. Such communities are generally sustained by their members' voluntary participation to generate content [Blanchard 1998]. This is usually indicated through posting and responding to messages and other electronic media that have been shared in the VC. Living in a community (e.g., registered as a member) does not guarantee participation, as demonstrated in both physical and virtual contexts [Preece 2004]. The members must be active enough to make the VC worth joining. Thus, a key challenge for most VCs is to keep on-going participation of their members [Blanchard 1998].

Virtual communities consist of users and "social objects" representing the intermediations of users' interactions.

In the following sections, we describe the different users' characteristics, such as user explicit profile, their activities, their connections, and the importance of trust in the case of interactions between users in the case of virtual communities in social platforms.

2.1.2 Main Pillars of Social Platforms

In social platforms, the virtual community is formed by their (i) Users and (ii) so called Social Objects representing the intermediations, topics of users' interactions (2.1). Social objects are commonly called User Generated Content in the case of the Web 2.0. In this case, they can be:

- Resources (representing shared objects, such as photos, videos, web pages, but can also be a physical object having a digital identity).

- Annotations (social content that describe a resource, such as social tags, microposts etc.)

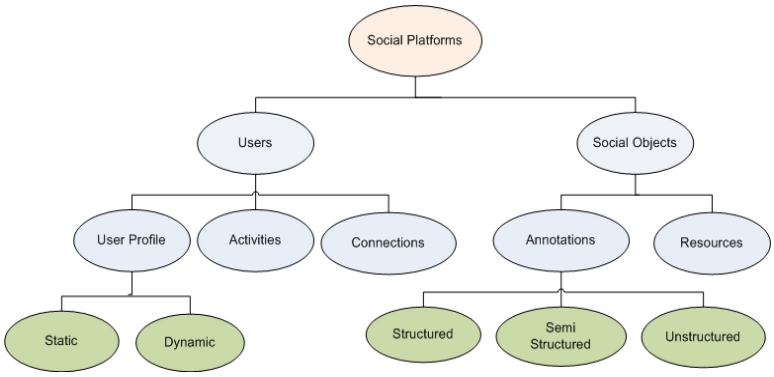


Figure 2.1: Taxonomy of Social Platforms: main pillars and corresponding dimensions

2.1.3 Users

Any social platform is composed of a set of users. In a given social platform, a user is characterized by a profile, activities and connections. First, we describe the different characteristics of the users, such as the user profile, activities, connections, and we inform on the importance of trust in social platforms. Then, we propose a concrete definition of a social object and study the different types of annotations on such objects.

2.1.3.1 The User Profile

The profile generally includes static personal information, such as the name, email and address, as well as more dynamic information about the interests and information needs of the user. The role of the user profile is essential in online communities. Generally user profiles are different from one application to another, as users present themselves differently, based on the targeted population of the given application (which are sometimes very specific). Thus, a user profile in a social networking system where the user has mostly friend connections will probably include more information about the hobbies and social activities. In a system targeted for professional networking, the profile will include information about the professional interests of the user, such as technical skills, jobs and future plans. More generally, a user can be a member of multiple social platforms and have different identity facets in each of them. An identity facet is a subset of the user's profile, targeted to a well-defined community.

2.1.3.2 Activities

Another dimension of users is represented by the activities they perform in the social platform. This includes content sharing, media uploading and content description (such

as photo tagging). In order to better understand the behavior of users in online communities, it is necessary to take into account some statistics about their activities and the amount of time spent performing them. According to the "Edison Arbitron Internet and Multimedia Study" [2] (year 2010), based on 2000 telephone interviews conducted in February 2010 in the United States, 48% of the interviewees possess a public profile in at least one social platform. This was 34% in 2009, and only 24% in 2008. The increase seems to accelerate as more and more applications are available.

Another important observation is that social network usage becomes a daily habit, as 30 of interviewees access their online accounts several times a day. Initially, the use of such application was considered only fun and meant for young people, but now the report clearly indicates a significant modification in the perception of such activities. Indeed, it has switched to an activity of socialization and social-awareness, as 25% of users are 18-25 years old and 23% between 25 - 34. The most widespread shared content on social networking websites are status messages, i.e. short content users post to express their current activity, interests or mood or to share a resource they consider interesting for their community. 72% of frequent social networkers post such a message frequently and 55% of the less avid users. 35% of users update their status several times a day.

2.1.3.3 Social Connections

Finally, the third dimension of users is represented by the social connections they establish with others in the platform . Users in these platforms are generally connected to different communities, belonging to different social spheres (e.g. friends, family, coworkers - Figure 2.2). Two major categories of connections can be distinguished in Social Platforms:

- Undirected Connections. This category of connections refers to the situation where a mutual agreement is necessary in order to establish the social tie. A good example is the case of friendships in social networking sites. The objective of such connections is to provide people with some kind of social awareness about the activities, mood and plans or their friends. This can be achieved by visiting their profile or by receiving notifications if there is an update in their social awareness stream.

- Directed Connections, a more recent form of establishing social ties. It means the fact that a user may want to receive notification from another, but not vice versa. This kind of connection is targeted to follow one's activity and not to express a friendship in real life.

[2] Edison Study on Social Network Usage - `http://www.edisonresearch.com/home/archives/2010/06/the_social_habit_frequent_social_networkers_in_america.php` - visited July 2010

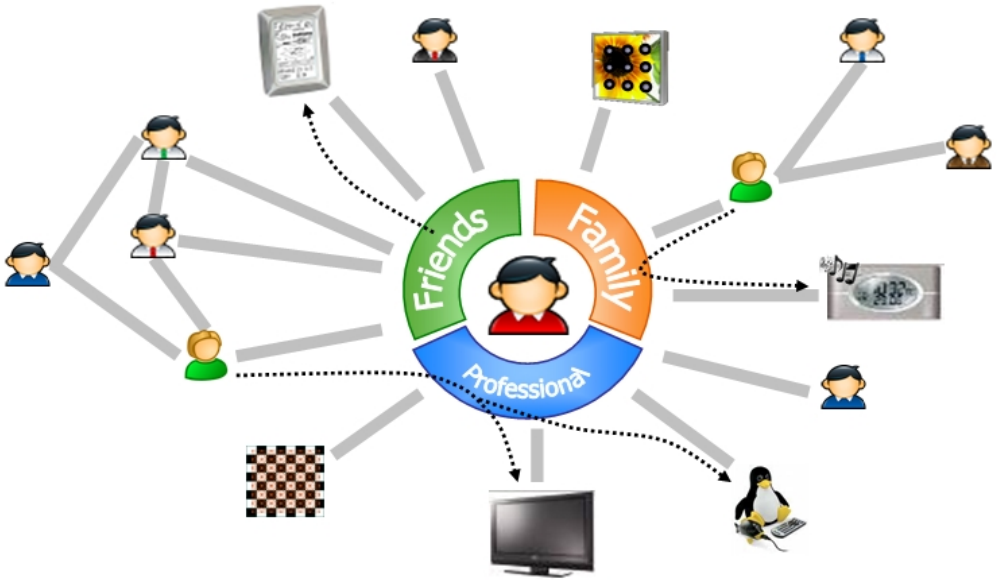


Figure 2.2: Visibility of a user in multiple social spheres

2.1.3.4 Trust

Another important users' characteristic is related to trust. Indeed, the different applications on social content sites allow users to be closer to their communities and to be aware of peers' activities and opinions. This brings new dimensions to trust and allows users to have higher confidence in the recommendations, suggestions and sentiment of friends. This tendency is also confirmed by the Socialnomics report [3] as friends' recommendations are much more trusted than recommendations from brand websites, consumer forums, television or newspapers. The increase of trust in this category of recommendations is due to the consumer generated content and the reliance on word of mouth (i.e. propagation of conversations in a community) in the decision making process, coming either from friends or followers or from people that are just participants in the community.

In conclusion, a user can be defined as an entity in online communities with a set of corresponding community-dependent faceted profiles, activities and connections, regulated by communication processes, such as trust, that influence their participation in the community.

[3] SocialNomics Report - `http://socialnomics.net` - visited September 2010

2.1.4 Social Objects

As mentioned before, shared content influences interactions between users. This observation gave birth to the object-centered sociality principle. An object is a common interest focal point, the "reason why people affiliate with each specific other and not just anyone" [4]. An object has a concrete and perceptible, physical and/or numeric, manifestation. It is a coherent indivisible whole, which triggers specific activities. Some objects are the source of conversational interactions and keepers of collective attention. They constitute a conversation support. In our actual digital context objects are mainly multimedia ones as articles (Wordpress, Wikipedia), videos (Youtube, Dailymotion), pictures (Flickr, Picasa). Annotations produced by users are used to describe the context or semantics of such artifacts.

They can be divided into:

- structured (i.e. semantic annotations, also called concepts)
- semi-structured (i.e. social tags)
- unstructured (i.e. free text, also called social awareness streams or status updates)

The manipulation of objects involves tasks such as description, retrieval, reuse, presentation and search. All these tasks need a layer of prior knowledge, which is represented by the annotations.

In the case of automatic annotation, the system automatically extracts features from the object (e.g. relevant descriptors for an image, keywords from a textual document etc.) and uses them as annotations. In the case of semi-automatic annotation, the system generally extracts the annotations from the resource, but asks the user to validate them. In the case of manual annotations, the user's cognitive capacity to interpret the meaning of an object is leveraged. In this case, the user has generally two possibilities for annotation: (i) the system gives the user complete freedom in choosing the term they intend to use in the annotation (the case of social tagging and free-text annotations), (ii) the system uses a vocabulary of terms and the user can choose a term from the vocabulary for the annotation (the case of semantic annotations and also some cases of social tagging). This second option gives users less freedom, but allows having a stable, convergent vocabulary that allows a better way of retrieving documents, as the description of resources will not suffer from synonyms, spelling errors or discrepancies in granularity. Also, this background vocabulary structure can be further used for the computation of similarities between annotated resources. In the following, we focus our attention on the annotations shared in social platforms.

Annotations may be either structured, semi-structured or unstructured:

[4] Engestrom, J. "Why some social network services work and others don't" - http://www.zengestrom.com/blog/2005/04/why_some_social.html (2005)

1. *Structured Annotations.* In this case, the terms employed in the annotation are regulated by a common domain vocabulary that must be used by the members of the system. These types of annotations are currently not used in the majority of social platforms because a domain vocabulary containing the necessary terms for the annotations is needed. Although such an approach has many advantages (e.g. absence of synonyms, absence of differences in pronunciation), this is not the natural way to describe resources in web 2.0 platforms, as the domain is not well-defined and, therefore, it is very difficult to build such vocabularies and to establish a consensus for each term used. At the same time, the use of semantic annotations would be cumbersome for people, as it is time-consuming and requires additional cognitive effort to select concepts from existing domain ontologies. In addition, semantic annotations work well in systems where the domain is well-defined (e.g. a system for sharing knowledge about human genes [Yeh 2003]), but in social platforms this is not the case, as the shared content is generally very heterogeneous, as people can discuss without limits (i.e. covers multiple domains with no regularities and relations).

2. *Semi-Structured Annotations.* In contrast, semi-structured annotations, such as social tags, are widely used in social platforms for photo tagging and bookmarking (e.g. the annotation of a web page). These annotations are generally freely selected keywords without a vocabulary in the background. However, we consider them to be semi-structured, as they represent an intermediate approach between semantic annotations (i.e. annotations that are based on concepts from domain ontologies) and free-text annotations. Besides, such collections of tags converge to a structured data organization, called a folksonomy [Gruber 2005]. This consists of a set of users, a set of free-form keywords (called tags), a set of resources, and connections between them. As folksonomies are large-scale bodies of lightweight annotations provided by humans, they are becoming more and more interesting for research communities which focus on extracting machine-processable semantic structures from them. These underlying data clouds of collaborative tagging systems enable Internet users to annotate or search for resources using custom labels instead of being restricted by pre-defined navigational or conceptual hierarchies (e.g. ontologies).

3. *Unstructured Annotations.* Finally, a more recent form of annotations is represented by free text annotations, also called social awareness streams, composed of status updates or microposts[Naaman 2010a]. This can be found in the majority of social networks and microblogging systems and primarily consists of free texts in the form of short messages describing a resource, a finding, an impression, a feeling, a recent activity, mood or future plan. The limitations of this practice from the viewpoint of information retrieval and knowledge management are sim-

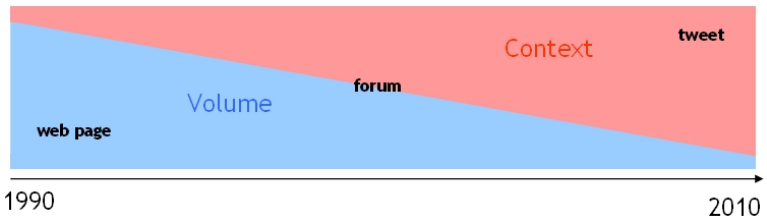


Figure 2.3: Evolution of content production on the Internet: from documents to microposts

ilar to that of social tagging, as users have complete freedom in the formulation of these messages. It is important to mention that in social awareness streams, the produced content often contains the described resource itself, in the form of an integrated hyperlink. A common practice is either to express an opinion about the resource (e.g. web page) or to provide its short summary for the community.

Since internet took over usenet as the main computer-based means of communication, it has gone through several stages: read-only web, with large pieces of information close to magazine article size; read-write web, or web 2.0, with forums mimicking usenet, exchanging pieces of information up to half-a-page in size; blogging, close to the web page model but with a shift in authorship towards the general public; and micro-blogging, based on very short messages (140 characters on Twitter). This shift from large, authoritative information to very short and amateur information is contemporary with the mobility evolution, with the more user-friendly web-enabled devices (e.g. the iPhone) emphasizing a particular factor: the context in which information is written. This has blurred the distinction between information and messaging, as all information on Twitter is in fact a message to followers, and all messages may be shared, thus creating information. Events and documentation on the contrary are becoming more distinct: in the traditional newspaper information model, documentation is delivered with events in a single article; in the Twitter-driven model, events are tweets, and the user is meant to seek information in more reliable and static sources, such as Wikipedia. An example of such as shift is the growing use of Twitter in the scientific community (studied in this article), contrasting strongly with the process of peer-reviewed publication.

An interesting issue about such free-form posts in social platforms is its short size, which emerged as a simple, convenient way to communicate about activities or share findings. The size limitation of such posts, defined by the majority of such platforms is mostly due to the fact that users can in this way follow hundreds of friends in real-time, without an important time investment. Also, this light-

weight form of communication enables users to broadcast opinions, activities and status [Java 2007] [Naaman 2010b].

The same applies to the composition of such posts, where common practices emerged as new means to better identify posts relevant to a specific event, also called "hashtags", or common ways to synthesize an information, such as including the source web page or reducing the amount of stop words in order to gain place for the informative terms (keywords and named entities). These practices largely depend also on the targeted user community, which can vary from a small family to the world at large. Also given its short size, such microposts are often called "social signals" [Mendes 2010b], and users of such systems "social sensors" [Sakaki 2010], as they can be useful to detect important events in a given location, such as an earthquake.

2.2 Classification and Discussion of Social Platforms

The so-called "Web 2.0" has introduced new freedom for users in their relation with the Web thanks to new online communication processes. Thanks to these artifacts, the user is now able to easily create content and make it available for the world at large, to annotate or add comments or to rate existing content. Some users spend a considerable amount of time on their social networks to exchange information within their social communities. An outcome of such technologies is the fact that now people are organized in communities around specific topics and interests. Interactions within the social network may be on a widely disparate and frequently unexpected range of topics, but they remain interesting to users because of the social relevance of the sources.

Based on the type of shared content and on the type of the resource it describes, we consider the following classification of social platforms:

- ***Social Networking Platforms - SNP*** were at the origin of the widespread acceptance of social networking on the Web, and drove the trend to develop social data. These platforms offer the basic form of tools to connect peers and to communicate, comment, share, rate, etc. resources and people. Today, social networking is the most popular activity in the Web.

- ***Social Publishing Platforms - SPP*** share some of the same functions, but with a focus on publishing content. This category of platforms is also called microblogging. This includes blogs, microblogs, as well as other shared content such as bookmarks and tags.

- ***Social Aggregation Platforms - SAP*** exist to augment one or more existing social networking applications, usually by providing a common interface to browse and/or publish to several underlying social networking sites. These social

SNP	SPP	SAP	SSP	SMP	SGP
Facebook	Twitter	FriendBinder	Aardvark	Wildfire	Aloqa
Myspace	Delicious	FriendFeed	Oneriot	Hypios	Foursquare
Secondlife	Twine	Silentale	Sysomos	Gravity	
LinkedIn	Yammer	Jodange	Swotti	Contexa	
Netlog	Posterous	Spokeo	Sproose	43 Things	
Hi5	Vox	Minggl	Moodviews		
Couchsurfing	StumbleUpon	Yonoo	Eurekster		
Habbo					
Tribe					

Table 2.1: Panorama of popular social platforms in each category

aggregators can differentiate themselves with additional functionality, ergonomy or analysis of the underlying data.

- ***Social Search Platforms - SSP*** are defined as search engines specifically dedicated to social data, including social content (such as shared media and comments) and social relations (such as finding a person in a network with certain characteristics or expertise). This means that these tools index social content and offer a means to the users to search that content, similarly to what Google does with the content of online web pages.

- ***Social Marketing Platforms - SMP*** are applications that exploit social data for viral marketing purposes. These are mainly companies created to reinforce the business model of social networking sites and make revenue from the huge amounts of data they have.

- ***Social Geolocation Platforms - SGP*** are platforms that exploit the mobility of people and their devices and enable them to declare their presence and activities to their social relations. These platforms offer the means to comment different objects associated to a well-defined physical location and to engage in a conversation with people that are present or connected to the given location.

Table 2.1 shows a panorama of the most popular online social platforms for each previously mentioned category.

A detailed description of a subset of the platforms mentioned in this work is given in the Appendix, section 7.3. A comparative analysis of these platforms with regards to a set of criteria is also included. This analysis helps understanding current trends in the design and implementation of these platforms.

At this stage, we can distinguish two main perspectives in the social web for the analysis of social platforms: (i) application perspective and (ii) research perspective.

From the application perspective, the objective is mainly to make the user's life as easy as possible in terms of communicating with peers, visualizing information in a crowd, aggregate different data flows, ergonomic interfaces, etc. It should be noted that at this level we mainly care about the front-end layer, thus about the end user, and *social information is considered in its native form*.

From the research perspective, the objective may be declined into two sub-objectives: (i) exploit social data for end-user's added value services and (ii) exploit social data for services providers. The research in this area intends also to improve the interaction between the user and the social platforms (data) but aims also to go beyond this by mainly analyzing the social interactions and understand what can be generated from those interactions. It is well established that the research community has focused mainly on the structural analysis of social interactions. Examples of the structural analysis include the calculation of key people, influencers, communities, etc. This information may be useful for the end user but is exploited mainly by services providers for, e.g., better marketing strategies.

It is clear that most of the social platforms consider the applicative dimension. This is true since most of them are targeted to the end-user. The most exception can be found certainly in some social marketing platforms. This can be explained by the usage of these platforms which is mainly targeted to services providers and not to end-users. Most of the platforms having a strong research investment consider the applicative dimension but in less priority than, e.g., the social networking platforms or social publishing platforms.

The majority of the studied social platforms consider an analysis of the structure of the underlying social network. Although the analysis is in a basic form, it is generally useful for the user. The analysis of the structure is generally done for suggesting new friends, identifying key-player, a community, etc. There is an increasing investment in this area from the different social platforms especially in the social marketing platforms. This can be explained by the need of very representative individuals in the network who need to be targeted while optimizing the cost of, e.g., a marketing campaign.

The other interesting dimension which started to attract research communities as well as industries is the content consideration. In fact, there is a growing interest in the exchanged content in social platforms. This can help in a better understanding of users' expectations. From the pragmatic perspective, the content may also help in solving some research problems related to the social networks structure understanding. The only platforms focusing on this dimension are platforms involving a heavy research investment and which build an additional layer of services for the user/service provider like search and marketing.

The most important issue for social platforms providers is certainly how to make revenue from their "social capital". From this perspective, most of the providers still rely on the traditional advertising techniques. There is no major innovation in the use

of the advertising technique also. This joins the issue of business models related to these platforms which is still problematic. In fact, except the social marketing platforms who build interesting models (this is normal since they deal directly with services providers and not with end-users), the other platforms need to be innovative in this area. The richness of social platforms is resides certainly in the number of its users. However, it is very difficult for all the platforms to concretize this richness for the moment. Many efforts are ongoing in this area.

From the mobility management perspective and even if there are still platforms which do not consider this aspect, some platforms innovate by either creating specific applications for the mobile to run their services, or by introducing new usages and even making it the only way to use the offered service. There is an increasing interest in the mobility management from the social platforms perspective due to, e.g., the huge time users spend using their mobile devices compared to their fixed devices like desktop computers.

Another emerging issue in the social platforms area is the consideration of real-time. The real-time issue relates to the recovery of information and their presentation to the user when they occur. This is built on the assumption that the most relevant information to the user on the social area is the most recent one. Measures are presented to consider this dimension especially in social search and real-time social platforms. The micro-blogging platforms, such as Twitter, are also considering heavily this dimension event is there is no considerable analysis on the data. It is clear that future applications need to consider in a form or another the real-time dimension because on mainly the heavy frequentation of users and the huge quantities of data which are produced.

One of the most important issues related to social platforms and to which end-users are becoming more and more aware, is privacy. Privacy has many aspects in social networks spanning from data access to information disclosure impact on real life. Generally speaking, all social platforms provide some ways to ensure privacy of users' data. The tools offered for this aspect are mainly based on traditional data access methods as defined in, e.g., databases. Thus, the user may define only the access or no access to a resource. There is no major innovation in the studied social platforms. The reason is that these platforms suppose that the privacy they offer is enough and the important issue resides in the "social" functionalities these platforms offer. As pointed out before, users are becoming more and more aware the impact their privacy on social platforms may have and are becoming very strict regarding this issue. This explains certainly the huge amount of ongoing research work in this area.

2.3 User Participation in Social Platforms

In the last part of our analysis of social platforms, we consider the issue of motivation, that concerns the users. In other words, we will try to summarize the main motivations

of users to share content in a social platform. This issue has already been investigated by several studies and we will only provide a short summary of them. [Naaman 2010a] outlines the following user motivation for content sharing in a social platform and provides corresponding abbreviations:

- Information sharing (IS): sharing mainly interesting web pages with the community
- Self promotion (SP): sharing information about personal projects, blogs or competencies
- Opinions/Complaints (OC): sharing opinions about products, technologies etc. or complaints about the same objects
- Statements and random thoughts (RT): sharing mainly feelings or current states of mind
- Me now (ME): sharing information about current activity
- Question to followers (QF): sharing an information need with the community
- Presence maintenance (PM): sharing information about current location
- Anecdote me (AM): sharing mainly stories about curious happenings, events that could be interesting to the community
- Anecdote others (AO): sharing mainly stories about curious happenings related to others

This statistical study shows that most messages are in the ME category (more than 41%) of the sampled dataset in Twitter. Other important categories of messages are the RT (25%), OC (25%) and IS (21%). Few questions are shared (only about 5%) which shows the fact as such content is generally difficult to identify in the flow of activity notifications of the followers. The fact that users share lots of content in these four categories shows that they may be an interesting resource to extract information in order to identify e.g. expertise.

A previous study worth mentioning in this category is that of [Java 2007], which confirms the fact that most posts on Twitter talk about daily routine or what people are currently doing. This is the largest and most common use of Twitter. Also, many users report latest news or comment about current events on Twitter. Some automated users or agents post updates like weather reports and new stories from RSS feeds. This is an interesting application of Twitter that has evolved due to easy access to the developer API.

In [Burke 2009] a study has been conducted on Facebook with the objective to identify content sharing behavior of newcomers in the platform. The most interesting findings of this study are the following:

- newcomers who see their friends contributing go on to share more content themselves
- For newcomers who are initially inclined to contribute, receiving feedback and having a wide audience are also predictors of increased sharing
- Newcomers whose initial content is distributed widely will go on to contribute more content
- Newcomers who are singled out in content will contribute more content

Users' experience in social networking sites is primarily a function of the content their friends contribute. If a user's friends post photos, compose blog entries, or exchange public messages on each other's walls, she can consume continually refreshing content. This provides an incentive for that user to continue logging in to the site, and might encourage her to contribute more content of her own.

A very important conclusion of this last study is related to *social learning*, e.g. learning from connections in the social platform. In other words, these results suggest design elements in social platforms which facilitate learning from friends, singling out, feedback, and content distribution can help increase the level of engagement for new users, leading to further content contributions and an overall better user experience.

2.4 Conclusion on the Pillars of Social Platforms

The objective of this chapter was the identification of the main constituents of social platforms and the underlying virtual communities providing a common environment for their conceptual modeling. In this context, we defined two main pillars for such systems, the users and social objects (also called user generated content) which are shared by users. Users of such platforms have generally a (i) basic static profile, composed of their personal information, such as name, location, age; (ii) activities such as photo tagging, content sharing and (iii) connections with people from various social spheres. Content they share is generally short and unstructured, as some systems even limit the size of posts (e.g. 140 characters in Twitter). This is also one of the main reason of their success.

Regarding the issue of content sharing, the main novelty offered to users by the social web is the complete freedom in sharing content and exchanging information with peers. This results in dynamic and evolving online communities. Social objects shared in such platforms may be both physical (e.g. a place, such as a train station) or digital

resources, such as a photo or a web page), as well as different annotations shared about these entities. It is also important to note that the preferred type of interaction in such platforms occurs in the form of short, unstructured messages, which allows users to share frequently information with lots of contextual metadata. In order to keep this sharing activity on-going, sociological and technical challenges need to be addressed with regards to an appropriate social information management strategy in these platforms. More specifically, the issue of how to better structure this huge amount of shared content in order to further improve on-going participation of members and reduce the knowledge latency needs to be discussed. For this reason, the next chapter introduces advances in several related fields: social information management, social network analysis, user modeling, semantic metadata management and privacy management in order to have a broad image of the state-of-the-art with regards to this problem.

Chapter 3

Social Information Management and Social Search

Contents

Before presenting our approach on information discovery and management in social content sites, this chapter will review existing work in the corresponding scientific fields and revisit related work in social information management and discovery. More specifically, we first identify the main scientific fields that contributed to the area of information processing and management in social platforms and that are strongly related to the two previously defined pillars. In a second time, we will target more closely frameworks that perform tasks such as social information discovery and organization and social search.

As pointed out in our analysis of the anatomy of Social Platforms, social platforms are composed of users and different social objects (annotations and resources) that regulate the dynamics and evolution of the underlying community. In the previous classification of these artifacts, we concluded that users are the main actors of such systems and are represented by their profile, activities and social connections. In the case of social objects, the two main categories are the annotations and the described resources (the Social Web community labels it also as User Generated Content (UGC)). Depending on the type of the platform, such annotations can be structured, semi-structured and unstructured. Clearly, the two last ones are the most frequent, as they offer the user complete freedom in the selection of keywords and entities used in the formulation of messages. From the system's and services provider's perspective, this freedom is however less compelling, as it yields difficulties in case of the efficient management of the content, such as their retrieval, extraction and classification of social data.

We consider the following fields for the review of social information management techniques in the case of the main pillars of Social Platforms: (i) Social Network Analysis, (ii) Semantic Metadata Management (i.e. conceptual models that capture social content) and (iii) User Modeling. Finally, we consider existing approaches for (iv) Privacy Management, a cross-field, as this is fundamental in systems that allow content sharing and some kind of social interaction between users. We focus in the last part on the specific area of Social Search frameworks.

We identify the following correspondence between the pillars, their dimensions and previously mentioned scientific fields (Figure 3.1):

- *Social Network Analysis.* In the case of social platforms, the user expresses his/her interests in the form of annotations (e.g. social tags or status updates). Therefore, we introduce current advances in Social Content Analysis, a field that provides tools for extracting knowledge from annotations or the understanding of the structural organization of virtual communities. This field is directly connected to the pillar social objects and more specifically, the annotations and connections. We introduce this field with a quick overview of the more traditional Social Network Analysis domain.

- *Semantic Metadata Management.* Contributions to the field of Semantic Metadata Management, a subfield of the more general Semantic Web, will help in better understanding from an upper-view the general structure of online communities and the underlying relationships between users and social objects. This field is related to the connections between the user and shared annotations.

- *User Modeling.* Advances in the User Modeling field allows us to understand the main strategies for capturing users' interests, preferences and storing this

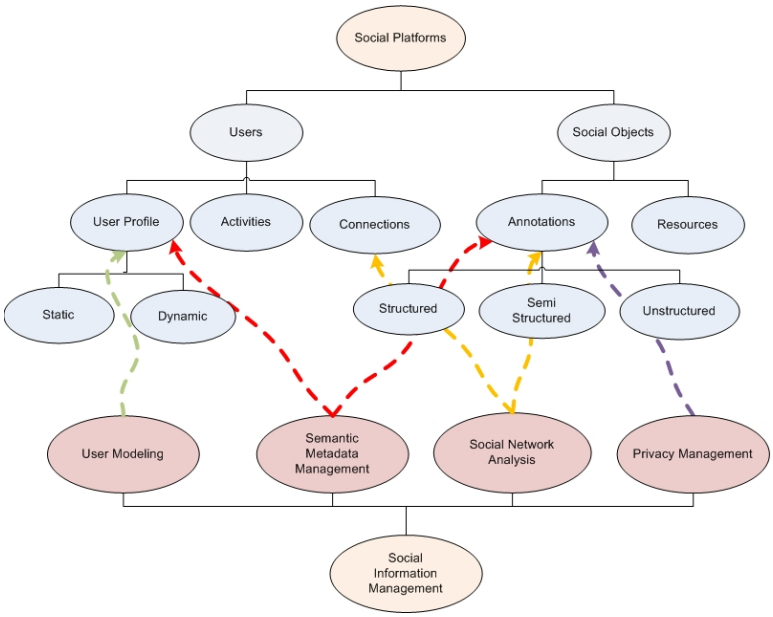


Figure 3.1: correspondence of fields to the pillars and their dimensions

information in a profile. This field is related to the user, the main pillar of Social Platforms.

- *Privacy Management.* In this specific case, the review of some fundamental approaches for user privacy management is also necessary, as it is a key component of successful recommendation strategies in a Social Platform, as shown by [Binder 2009a]. This field is related to the social objects and more specifically, annotations. Also, privacy management is necessary for us for the conception and implementation of a model that takes into account trust and intimacy in our approach.

3.1 Social Network Analysis

As introduced before, the underlying human organization in Social Platforms is in the form of online (virtual) communities, physically structured into social networks. Formally, a social network is a collection of vertices and edges connecting them. In traditional social networks, vertices are the users and edges the friend relationships between users. The analysis of such a network comprises different algorithms to understand its topology, user's behavior and the patterns and meaning of shared content.

Thus, a basic division of analysis strategies in social networks can be considered

(Figure 3.2):

- Structural social network analysis, based on the topology of the network (i.e. structural properties, like the degree of a vertice - e.g. number of incoming and outgoing connections -)
- Semantic social network analysis, based on the analysis of the exchanged and shared content between peers or online communities

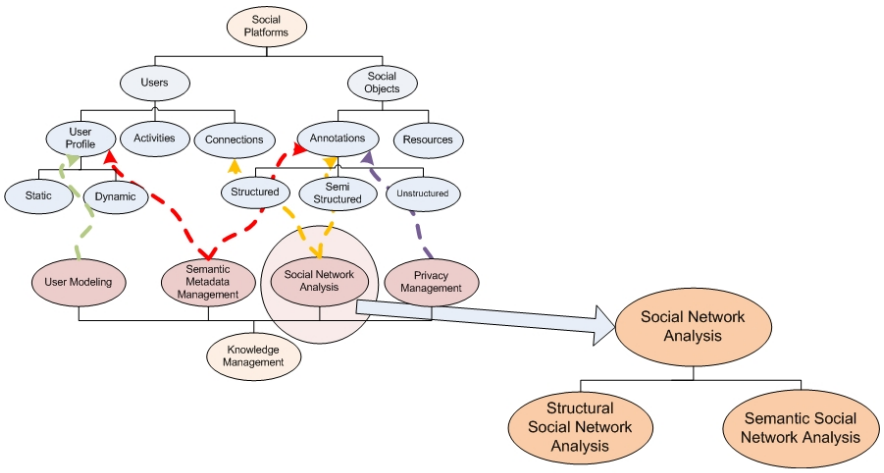


Figure 3.2: Main division of Social Network Analysis methodologies: structural and semantic analysis

3.1.1 Structural Social Network Analysis

Structural social network analysis considers the statistical distribution of vertices and links in the network in order to extract communities of strongly connected users, understand the mathematical (statistical or probabilistic) functions that govern the distribution of these connections and predict the evolution of the network.

The two most important scientific results of this analysis strategy have been achieved in the areas of: (i) Community Extraction, (ii) Topology Analysis and (iii) Link Prediction.

More concretely, we can cite the two following proved hypothesis:

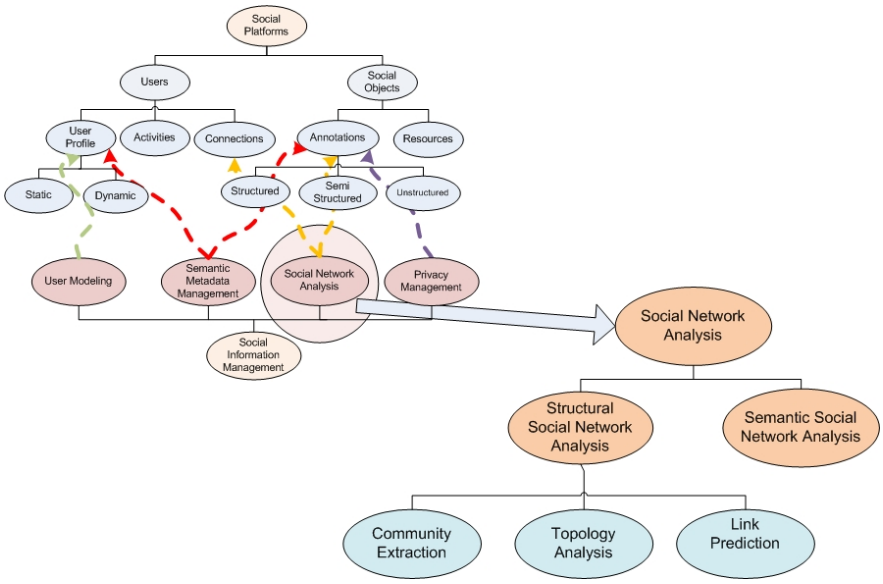


Figure 3.3: Main subfields of Structural Social Network Analysis

- The understanding of the fact that the distribution of connections in the network follows the scale-free power-law function [Newman 2006].
- The discovery of the 6 degrees of separation principle (small worlds) [Leskovec 2008].

A common property of many large networks is that the vertex connectivities follow a scale-free power-law distribution [Newman 2006]. This feature was found to be a consequence of two generic mechanisms: (i) networks expand continuously by the addition of new vertices, and (ii) new vertices attach preferentially to existing ones that are already well connected. Such power-law distributions are known in many different articulations, such as the distribution of the wealth in a population (Pareto's law [del Castillo 2009]), that of the active words in English language (Zipf's law [i Cancho 2001]) and in many biological systems also. As stated by [Barabási 1999], the sociological basis of the power-law distribution is the *"preferential attachment"*, meaning that popularity is attractive:

> "These examples indicate that the probability with which a new vertex connects to the existing vertices is not uniform; there is a higher probability that it will be linked to a vertex that already has a large number of connections. Because of the preferential attachment, a vertex that acquires more connections than another one will increase its connectivity at a higher rate, thus, an initial difference in the connectivity between two vertices will increase further as the network grows."

The probability with which a new vertex connects to the existing vertices is not uniform, as there is a higher probability that it will be linked to a vertex that already has a large number of connections. A vertex that acquires more connections than another one will increase its connectivity at a higher rate, thus, an initial difference in the connectivity between two vertices will increase further as the network grows.

Concerning the 6 degrees of separation, this is a consequence of the fact that the network is composed of small strongly connected communities (small worlds) and vertices that connect these communities (key players). As a result, the theory states that it is possible to connect any two vertices in the network with a path composed of, at most, 6 vertices. First stated by Karinthy in 1929 and further confirmed by the famous Milgram experiment in 1967, as well as by works by Leskovec and Horvitz [Leskovec 2008] on online social networks allowed refining the supposed minimal path between any two vertices to an average of 6.5.

A community in a social network is generally defined as a set of vertices that are more highly connected to each other, than to the rest of the vertices [Lancichinetti 2008] [Palla 2005].

Such communities are interconnected by nodes (i.e. people), called key players, i.e. nodes with high betweenness centrality. Betweenness centrality of a node reflects to what extent it is between other vertices. Nodes with high betweenness centrality are very important in the network as other vertices are connected with each other mainly through them [Kajdanowicz 2010].

Structural social network analysis applications are mainly used for the extraction of such communities from social networks (e.g. clustering of the network using centrality indices to find community boundaries - [Girvan 2002] [Newman 2006]), the identification of key players [Borgatti 2006] and the prediction of the evolution of the network (e.g. link prediction [Leroy 2010] [Lu 2009]). This latter problem attracts lots of attention lately, as there is an increasing number of applications in telecommunications and defense industry (detection of potentially suspicious communication patterns [Dasgupta 2008], prediction of the evolution of criminal networks [Xu 2005] and the understanding of the propagation of viruses [Wang 2003]).

The analysis of the network structure is generally done to suggest new connections, identify a key-player or a community. There is an increasing investment in this area in different Social Platforms (e.g. friend recommendation in Facebook [1] or Twitter [2]), especially in the category of social marketing. This can be explained by the need for very representative individuals in the network who need to be targeted in order to optimize cost for e.g. a marketing campaign. A conclusion about structural analysis could be that this is more useful for researchers and service providers, but there are few compelling applications that demonstrate the practical usefulness directly for the end-

[1] Facebook Social networking site - `www.facebook.com` - visited January 2011

[2] Twitter Social Microblogginh site - `www.twitter.com` - visited January 2011

user. For this reason, the analysis of the shared content (e.g. interactions, exchanges) in the network has been introduced as a potential way to better understand user's needs and interests.

In the next section we review related work in Semantic Social Network analysis, focusing primarily on how semantic information can be extracted from social objects.

3.1.2 Semantic Social Network Analysis

A second strategy to analyse social networks, called semantic analysis, considers the shared content extracted from exchanges in the network instead of its topology. The main added-value of this strategy compared to the previous one is the fact that shared content represents much better the interests and information needs of an individual, than its social connections.

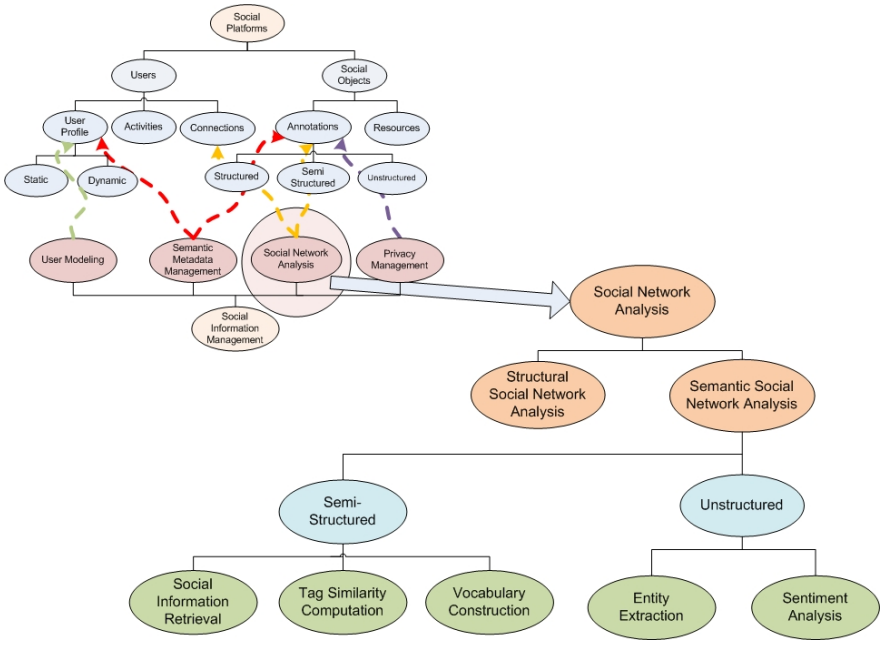


Figure 3.4: Semantic Social Network Analysis

Applications of semantic analysis include recommendation strategies that target either the preferences of an individual or a group of individuals, interest-based community extraction [Mika 2007a], entity extraction (i.e. the extraction of names, organizations, technology names from exchanges) for content indexation [Culotta 2004], content summarization (i.e. extracting a summary of exchanges in a social network or community) [Tseng 2007], sentiment analysis [Godbole 2007] and opinion mining [Domingos 2005].

The main difficulty in this analysis strategy is the fact that existing methods for semantic content management need to be adapted to the specificities of annotations in Social Platforms. As seen in Chapter 2.2, Section 2.1.2, three main types of annotations can be found in Social Platforms: (i) unstructured, (ii) semi-structured and (iii) structured. Since only very few Social Platforms (e.g. SMOB [3] -Semantic-MicrOBlogging- [Passant 2008c], Faviki [4] [Milicic 2008]) use the last one, in the following we focus only on the case of the management of semi-structured and unstructured annotations (Figure 3.4).

As explained before, the main reason for this is the fact that the current habit of users is to share content freely, without any vocabulary restrictions. A common form of semi-structured annotations in Social Platforms are *Social Tags* [Suchanek 2008]. These are keywords used to annotate mostly photos and the bookmarking of web pages. In the case of unstructured annotations, currently *Social Awareness Streams*, composed of *Status Updates* are the most widespread [5].

3.1.2.1 Analysis of Semi-Structured Annotations (Social Tags)

A first category of annotations in Social Platforms are semi-structured, also called social tags. Social bookmarking systems [6] have become extremely popular in recent years. Their underlying data structures, known as folksonomies [Mathes 2004], consists of a set of users, a set of free-form keywords (called tags), a set of resources, and a set of tag assignments, i.e., *(user - tag - resource)* triples. As folksonomies are large-scale bodies of lightweight annotations provided by humans, they are becoming more and more interesting for research communities that focus on extracting machine-processable semantic structures from them (we address this issue in Section 3.2). Collaborative tagging generally refers to the tagging of a collection of documents commonly accessible to a large group, rather than tagging contents located all over the Web, which is instead called social bookmarking. The main property of such systems is the fact that users tag documents with freely selected keywords, instead of using domain vocabularies.

Folksonomies contain peoples' structural knowledge about documents. A person's structural knowledge has been defined as the knowledge of how concepts in a domain are interrelated from the individual's point of view. According to [Mathes 2004], an important aspect of a folksonomy is that it is comprised of terms in a flat namespace: that is, there is no hierarchy, and no directly specified parent-child or sibling relationships between these terms. There are, however, automatically generated "related" tags, which cluster tags based on common URLs. This is unlike formal taxonomies and classification schemes where there are multiple kind of explicit relationships between terms.

[3] SMOB - `www.smob.me` - visited January 2011
[4] Faviki - `www.faviki.com` - visited January 2011
[5] concepts defined in Chapter 2.2 Section 2.1.2
[6] Delicious - `www.delicious.com` - visited April 2010

These relationships include functions like broader, narrower, as well as related terms. These folksonomies are simply the set of terms that a group of users tagged content with, they are not a predetermined set of classification terms or labels.

Folksonomies claim to have many advantages over controlled vocabularies or formal taxonomies. Tagging has lower costs because there is no complicated, hierarchically organized vocabulary to learn and adapt to its own one. Users simply create and apply tags. According to Wu et al., "Folksonomies are inherently open-ended and therefore respond quickly to changes and innovations in the way users categorize content" [311 2006]. Collaborative tagging is regarded as democratic metadata generation where metadata is generated by both the creators and consumers of the content.

Folksonomies can be divided into broad folksonomies, which allow different users to assign the same tag to the same resource, and narrow folksonomies, in which the same tag can be assigned to a resource only once [7].

The question of why folksonomies are successful has been the subject of several studies in the literature. An important argument for this is the fact that the feedback loop is tight [Mathes 2004], i.e. once the user assigns a tag to an item, the cluster of items with identical or similar tags can be immediately retrieved. This can help the user decide whether to keep the tag or change it to a similar or different one. The scope of such a cluster can be expanded by showing all items from all users in the system which are tagged with the same tag. By viewing the result set, the user can decide how to better adapt the tag to the group norm or to have better visibility in the community for the tagged resource. The issue of how to influence the group norm was also studied by Udell [8]. This tight feedback loop leads to a form of asymmetrical communication between users through metadata. The users of a system are negotiating the meaning of the terms in the folksonomy, whether purposefully or not, through their individual choices of tags to describe documents for themselves.

A folksonomy eases collaboration. Groups of users do not have to agree on a hierarchy of tags or detailed taxonomy, they only need to agree, in a general sense, on the "meaning" of a tag enough to label similar material with terms for there to be cooperation and shared value. Although this may require a change in vocabulary for some users, it is never forced, and as Udell discussed, the tight feedback loop provides incentives for this cooperation.

The main problems of social tagging systems include ambiguity, lack of synonymy and discrepancies in granularity [Golder 2005]. An ambiguous word, e.g. apple, may refer to the fruit or the computer company, and this in practice can make the user retrieve undesired results for a given query. Synonyms like lorry and truck, or the lack of consistency among users in choosing tags for similar resources, e.g., *nyc* and *new*

[7] http://www.personalinfocloud.com/2005/02/explaining - visited April 2010

[8] `http://www.infoworld.com/d/developer-world/collaborative-knowledge-gardening-020` - visited April 2010

york city, makes it impossible for the user to retrieve all the desired resources unless he/she knows all the possible variants of the tags that may have been used. Different levels of granularity in the tags may also be a problem: documents tagged java may be too specific for some users, but documents tagged programming may be too general for others.

Several attempts have been made to uncover the structure of this kind of data organization. Basic formal models of folksonomies include that of Mika [Mika 2007a] and Hotho et al. [Hotho 2006]. Mika proposes a model based on *tripartite hypergraphs*, while Hotho et al. on *triadic context* (term used in formal concept analysis). We present in the following the formal model of Mika, one of the most cited models in the literature for the representation of these structures.

As said before, a folksonomy is an association of users, annotations and resources. The corresponding three disjoint set of vertices are considered by Mika in the formal model: the set of actors (users) -A-, the set of concepts (tags) -C- and the set of resources -O- (e.g. photos, videos or web resources, like bookmarks, websites etc). Since in a social tagging system, users tag objects with concepts, ternary relations are created between the user, the concept and the object.

This resulting tripartite hypergraph can be transformed into several bipartite graphs, each having a very specific meaning, like AC - the graph that associates actors and concepts, CO - the graph that associates concepts and objects and AO, the graph that associates actors and resources.

Abel [Abel 2008a] investigates the benefits of additional semantics in folksonomy systems. Additional context can be provided to the tagging activity with an extension of the tripartite model, i.e. an association of the user, the tag and the tagged resource, that describes more precisely the particular tagging activity. For example, time stamp helps to categorize tags in a temporal manner, the mood the user had when tagging the resource would allow to qualify opinions expressed in a tag. Other information, like background knowledge about the user, would allow to have information about the reliability of the tagger. The GroupMe! folksonomy system is proposed, which is a new kind of resource sharing system for multimedia web resources. A first extension of previous models is the introduction of the term group, which is a finite set of related resources. The folksonomy model in GroupMe! can be thus formalized in the following manner (We note with F the folksonomy model): $F = (U, T, IR, G, Y)$, where U, T, R, G are finite sets that contain instances of users, tags, resources and groups. $IR = R \cup G$ is the union set of resources and the set of groups.

Wu et al. [Wu 2006] identify the key challenges in collaborative tagging systems. The three identified challenges are the following: (i) the identification of communities, i.e. groups of users with similar interests, (ii) preventing information overload by filtering out high quality documents and users (e.g. experts in a domain) and (iii) how to create scalable, navigable structures from folksonomies. Folksonomies are criticized to

have flaws that formal classification systems are designed to eliminate, including polysemy, words having multiple related meanings, and synonymy, multiple words having the same or similar meanings.

Information retrieval from folksonomies: Social Information Retrieval

In the previous section we have seen the general definition and structure of folksonomies, the data organization in social tagging systems. In this section we go further and review existing techniques of information management in folksonomies.

The biggest challenge in folksonomies is information retrieval, i.e. the question of how to efficiently rank items (e.g. tags, resources, users) for a given user query. In traditional Internet applications the search and navigation process serves two vital functions: retrieval and discovery. Retrieval incorporates the notion of navigating to a particular resource or a resource containing particular content. Discovery incorporates the notion of finding resources or content interesting but theretofore unknown to the user. The success of collaborative tagging is due in part to its ability to facilitate both these functions within a single user-centric environment. Reclaiming previously annotated resources is both simple and intuitive, as most collaborative tagging applications often present the user's tag in the interface. Selecting a tag displays all resources annotated by the user with that tag. Users searching for particular resources they have yet to annotate may select a relevant tag and browse resources annotated by other users. However, the discovery process can be much more complex. A user may browse the folksonomy, navigating through tags, resources, or even other users. Furthermore, the user may select one of the results of a query (i.e. tag, resource, or user) as the next query itself. This ability to navigate through the folksonomy is one reason for the popularity of collaborative tagging.

In order to provide efficient retrieval mechanisms, a formal model of folksonomies is required. There are several models in the literature, e.g. that of Mika [Mika 2007a] and Hotho et al. [Hotho 2006]. Mika proposes a model based on *tripartite hypergraphs*, while Hotho et al. on *triadic context* (term used in formal concept analysis).

Hotho et al. adapt the well-known PageRank algorithm in order to apply it on folksonomies, called *FolkRank*. The impossibility of applying *PageRank* has its origins in the fact that a folksonomy is different from the web graph (undirected triadic hyperedges instead of directed binary edges). By modifying the weights for a given tag, FolkRank can compute a ranked list of relevant tags.

The original formulation of PageRank [Brin 1998] reflects the idea that a page is important if there are many pages linking to it, and if those pages are important themselves (recursive aspect of importance). The distribution of weights can thus be described as the fixed point of a weight passing scheme on the web graph. This idea was extended in a similar fashion to bipartite subgraphs of the web in HITS [Kleinberg 1999] and to

n-ary directed graphs [Xi 2004]). The same underlying principle is employed for the ranking scheme in folksonomies. The basic notion is that a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. Such a ranking schema can help the emergence of a common vocabulary in collaborative tagging systems, by recommending to the user tags that have a bigger visibility in the community and that is also semantically close to the user-defined tag.

Abel et al. [Abel 2008b] perform an in-depth analysis of ranking algorithms specially designed for folksonomies: FolkRank, SocialSimRank [**?**], and SocialPageRank and adapts them to the GroupMe! social bookmarking system, where an additional dimension is added to folksonomies, i.e. groups of resources.

Gemmel et al. [Gemmell 2008] propose a method to personalize a user's experience within a folksonomy using unsupervised clustering of social tags as intermediaries between a query and and a set of items. Terms in the query are weighted based upon their affinities to particular clusters to help disambiguate queries.

Bao et al. [Bao 2007] propose different algorithms, such as SocialSimRank and SocialPageRank to optimize web search using social annotations. The undelying hypothesis of the proposed algorithms are the following: (i) social annotations about web pages are good summarizations of the given web page and can be used for efficient computation of similarity between a search query and a web page and (ii) the amount of annotations assigned to a web page is a good indication of its popularity.

Vocabulary Construction and Emergence of Semantics

In this section we present different approaches for extracting and constructing a hierarchical structure of tags in collaborative tagging systems. Recently, several papers proposed different approaches to construct conceptual hierarchies from tags collated from social Web sites. Mika [Mika 2007a] uses a graph-based approach to construct a network of related tags, projected from either a user-tag or object-tag association graphs. Although there is no evaluation of the induced broader/narrower relations, the work provides a good suggestion to infer them by using betweenness centrality and set theory. Other works apply clustering techniques to keywords expressed in tags, and use their co-occurrence statistics to produce conceptual hierarchies [Brooks 2006] [Zhou 2007]. In a variation of the clustering approach, Heymann [Heymann 2006] uses graph centrality in the similarity graph of tags. In particular, the tag with the highest centrality would be more abstract than that with a lower centrality; thus it should be merged to the hierarchy before the latter, to guarantee that more general node gets closer to the root node. Schmitz [Schmitz 2006] has applied a statistical subsumption model to induce hierarchical relations of tags.

Brooks et al. [Brooks 2006] argue that hierarchical structures which seems to match

that created by humans can in fact be inferred from existing tags and articles in collaborative tagging systems. This may imply that folksonomies and traditional structured representations are not so opposed after all, rather, tags are a first step in helping an author or reader to annotate her information. Automated techniques can then be applied to better categorize specific articles and relate them more effectively to other articles. The method used is agglomerative clustering and consists of the following steps: the comparison of each tag cluster to every other tag cluster, using the pairwise cosine similarity metric. Each article in cluster one is compared to each article in cluster two and the average of all measurements is computed. The two closest-similarity clusters from the list of tag clusters is removed and replaced with with a new abstract tag cluster, which contains all of the articles in each original cluster. This cluster is annotated with an abstract tag, which is the conjunction of the tags for each cluster.

This procedure is followed until there is a single global cluster that contains all of the articles. By recording the order in which clusters are grouped into progressively more abstract clusters, a tree that shows the similarity of tags can be constructed. Plangprasopchok et al. [Plangprasopchok 2009] proposes a different approach for constructing folksonomies from user-specified relations on Flickr [9] by statistically aggregating tags from different collections. This approach uses the shallow hierarchies created through the collection-set relations on Flickr. Authors argue that partial hierarchies are a good source information for generating folksonomies and propose a simple statistical approach to resolve hierarchical relation conflicts in the aggregation process.

Another approach for the extraction of hierarchical semantics from social annotations is proposed by Zhou et al. [Zhou 2007]. A probabilistic unsupervised method is proposed, called Deterministic Annealing. This method performs a top-down approach on the flat tag space, beginning with the root node containing all annotations and splitting it to obtain clusters with narrower semantics.

[Cattuto 2008] performs an analysis on a large-scale snapshot of the popular social bookmarking system Delicious [10]. To provide a semantic grounding of the folksonomy-based measures, tags of of delicious are mapped to synsets of WordNet [Markines 2009] and use the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, the similarity is measured by using both the taxonomic path length and a similarity measure by Jiang and Conrath [Jiang 1997] that has been validated through user studies and applications [Budanitsky 2006]. The use of taxonomic path lengths, in particular, allows to inspect the edge composition of paths leading from one tag to the corresponding related tags, and such a characterization proves to be especially insightful. Co-occurence is a measure that extracts from the folksonomy a graph for tags, where edges are weighted with the number of times they co-occur (i.e. tags on the same resource).

[9] Flickr Photo Sharing and Tagging System - `www.flickr.com` - visited July 2010
[10] www.delicious.org - visited April 2010

The results can be taken as indicators that the choice of an appropriate relatedness measure is able to yield valuable input for learning semantic term relationships from folksonomies, i.e. (i) synonym discovery, (ii) concept hierarchy extraction and (iii) the discovery of multi-word lexemes. The cosine similarity is clearly the measure to choose when one would like to discover synonyms. Cosine similarity delivers not only spelling variants but also terms that belong to the same WordNet synset. Both FolkRank and co-occurrence relatedness yields more general tags. This could be a proof that these measures provide valuable input for algorithms to extract taxonomic relationships between tags.

An important issue in the semantic analysis of content is the capacity to compute similarities between content items.

Tag Similarity Measures

In the following, we revisit the most well-known similarity measures known in the literature.

Statistical Similarity Measures

According to [Markines 2009], the most important statistical similarity measures used in Social Platforms are: (i) Matching, (ii) Overlap, (iii) Jaccard, (iv) Dice and (v) Cosine. In the following we consider X_i as a set of tags, that describe a given resource (e.g. a user), say x_i. Each tag has a weight, w_{x_i}. We note $|X| = \sum_y (w_{xy})$, the product of all elements in the set.

The previously mentioned similarity measures then take the following form:

- Matching Similarity:

$$\sigma(x_1, x_2) = \sum_y (w_{x_1} w_{x_2 y}) = |X_1 \bigcap X_2|$$

- Overlap:

$$\sigma(x_1, x_2) = \frac{|X_1 \bigcap X_2|}{min(|X_1|, |X_2|)}$$

- Jaccard:

$$\sigma(x_1, x_2) = \frac{|X_1 \bigcap X_2|}{|X_1 \bigcup X_2|}$$

- Dice Coefficient:

$$\sigma(x_1, x_2) = \frac{2|X_1 \bigcap X_2|}{|X_1|+|X_2|}$$

- Cosine Similarity:

$$\sigma(x_1, x_2) = \frac{X_1}{\sqrt{|X_1|}} \frac{X_2}{\sqrt{|X_2|}} = \frac{|X_1 \bigcup X_2|}{\sqrt{|X_1||X_2|}}$$

Semantics-Enabled Similarity Measures

This category of measures differs from the previous, as it employs a taxonomy (e.g. a vocabulary with relations between terms) to compute the distances between keywords. The most well known measures are the following: (i) Rada et al. [Mihalcea 2006], (ii) Wu et al. [Wu 1994], (iii) Chodorow et al. [Leacock 1998], (iv) Jiang et al. [Jiang 1997] and (v) Resnik et al. [Resnik 1999].

Rada et al. [Mihalcea 2006] is probably the simplest and most intuitive measure, as it counts the shortest path in the concept graph between the two concepts to be compared:

$\sigma(c_1, c_2) = Min(\varepsilon(X_1, X_2))$, where ε represents the number of edges between the two concepts.

[Wu 1994] improves the previous measure by introducing a normalization by taking into account the closest ancestor concept in the hierarchy (C_3(e.g. the concept that is closest to both). We further consider N_i, the number of nodes between C_i and C_3. N_3 represents the number of nodes from the ancestor concept to the root of the concept hierarchy.

$$\sigma(c_1, c_2) = \frac{2*N_3}{N_1+N_2+2*N_3}$$

Chodorow uses a similar technique, by taking into account the minimal path length between two concepts and the depth of the hierarchy. A particular category of distance measures is represented by Resnik et al. and Jiang et al., as they both take into account for each concept its probabilistic informativeness. This is a probabilistic measure from Information Theory, which attempts to evaluate the probability to find the given concept in a given context.

3.1.2.2 Analysis of Unstructured Annotations: Social Awareness Streams

The majority of Social Platforms consider a structural analysis of the underlying community. The other interesting dimension which has started to attract research communities as well as industries is the integration of semantic technologies, both on the modeling and content exploration level. There is a growing interest in the exchanged

content inside the social interactions in social networks. This can help to better understand users' expectations. From a pragmatic perspective, the content may also help in solving some research problems related to the understanding of social networks structure. The only platforms that focus on this dimension are platforms involving a heavy research investment and which build an additional layer of services for the user/service provider, like search and marketing. The most important issue for Social Platform providers is certainly how to make money from their "social capital". From this perspective, most of the providers still rely on traditional advertising techniques, with little or no major innovation. This also touches on the more general issue of business models related to these platforms, which is still an unsolved issue. In fact, except for social marketing platforms who build interesting models (since they deal directly with services providers and not with end-users), the other categories of platforms need to be innovative in this area.

The most common form of shared content in social networks in represented by the previously mentioned social awareness streams. Semantic analysis of social networks proposes different strategies to analyse such streams and to build different recommendation services that leverage knowledge extracted from it.

Lately, the analysis of such streams receives increased attention from different communities for two main reasons: (i) nowadays, such messages represent the most frequent form of interaction in Social Platforms and (ii) the understanding of such messages is a valuable source for building advanced recommendation strategies that have up-to-date information about the current needs and preoccupations of an individual or a community. An important observation is also the fact that shared content provides a richer representation of an online community at a given time, as it changes more rapidly than the topology (e.g. relations between people).

Similarly to the attempt to formally model folksonomies, in a recent work, [Wagner 2010] introduces the concept of *"Tweetonomies"*, a term syntaxically and conceptually similar to that of folksonomies. The main idea is the introduction of statistical measures on social awareness streams that allows to compare different streams. Inspired by the early work of [Mika 2007a], a Tweetonomy is defined as a ternary relation between users, messages and resources. Each element in this set is assigned a qualifier, which defines their role in the specific context (e.g. a user can have the role of author of a messages, but can also be mentioned in a message). The introduced measures for Tweetonomies attempt to capture the diversity of users, topics and vocabulary used (e.g. lexical diversity) with the help of simple statistical measures. With this formalization, it is possible to compare different social-awareness streams and measure their mutual diversity and overall quality.

The most important semantic analysis strategies that can be performed on social awareness streams to extract additional knowledge include: (i) Sentiment Analysis (i.e. also called Opinion Mining, is the extraction of the sentimental polarity of a message)

and (ii) Entity/Keyword Extraction (i.e. the extraction of named entities or keywords from a message and their disambiguation). In the case of social awareness streams, the main difficulty consists in the fact that there are few contextual cues to correctly disambiguate a named entity, as the messages are generally very short and completely unstructured (e.g. 140 characters allowed in Twitter).

In the following, we revisit related work in entity extraction and sentiment analysis.

Entity Extraction

Entity Extraction is the effort of extracting entities, such as people's names, technologies, places and institutions from text. The most important difficulty in entity extraction is the issue of name variations. Common types of name variations include: (i) lexical (e.g. organization, organization), (ii) orthographic (e.g. Rocky 2, Rocky II), (iii) structural (day of birth, birthday) and (iv) morphological (plural, singular variations, like mouse - mice). Also, a common problem is represented by ambiguous meanings (e.g. Calvin: theological or comics figure? Apple: company name or fruit?). Entity Extraction techniques generally perform a text-pre-processing that includes the splitting of text into sentence boundaries, tokenization and word stemming (e.g. Porter stemming [Porter 1980]).

As social awareness stream messages are generally short, it is important to note that issues like anaphora (an element in the text which depends for its reference on the reference of another element) and metonymy (a figure of speech consisting of the use of the name of one thing for that of another) are rare. After the pre-processing, several techniques can be applied to extract named entities [Nadeau 2007]: (i) lexicon-based (e.g. the comparison of keywords to terms in a lexicon), (ii) regular expressions (the definition of rules that specify the syntax of company names, people names, locations etc.), (iii) statistical classifier-based, like boundary window and sliding window token and (iv) finite state machines.

The advantage of lexical methods is the fact that often such vocabularies contain additional information about entities, such as the postal code for a location or the web page of a person. The clear disadvantage is certainly the fact that the matching of the keyword with terms in the vocabulary does not take into account the context of the keyword (e.g. Denver will be matched to a location even if it is a person's name). Another disadvantage is the fact that such vocabularies need continuous maintenance which is costly for organizations and people.

Regular expressions define common patterns for different kinds of entities. The advantage is that a learning mechanism can generalize from particular examples and choose the closest class for a given keyword. The disadvantage is similar to lexical based methods: context is generally not taken into account in the matching process. Also, the definition of such patterns is a tedious task. Classifiers like boundary window or

sliding window attempt to classify tokens in the text by taking into account the local context and learn features that describe different entity types. The main problem with such an approach is the necessity of training data that is annotated with the types of named entities. An intermediate approach is the use of Hidden Markov Models. Based on annotated data, such a model learns transition probabilities between elements in the text and predicts the most probable category of an entity.

A particular difficulty is the disambiguation of the extracted keywords and named entities.

In the following, we review the most well-known existing approaches in this field in order to have a clear vision of how they must be adapted to our specific context, which is provided by the underlying mechanisms that govern social platforms.

Disambiguation is a common problem in computer language processing. Different domains proposed solutions to resolve it (e.g. Machine Learning, Statistical Language Processing). In the Semantic Web domain, and more precisely in the context of our research, we have to match a keyword or an entity to the right Link Data concept that best approximates its meaning (e.g. Entity "Apple" in a message can refers to both concepts "Apple Inc." as a company or "Apple" as a fruit). Moreover, the short nature and lack of contextual cues make the ambiguous situations even more difficult.

The most interesting outcome of user generated tag is the ability to study user's interest. In [Li 2008] and [Golder 2006], the authors indicate that the tagging activities of a user carries interesting informations about his/her interest. [Kim 2003], proposed a user's interest hierarchy for defining user's interest. Their work suggests that text and phrases from users' bookmarks can be used to identify users' interest from general to specific.

The following most common approaches exist for disambiguation:

- Ontology based approach To address the aspect of lack of formal and explicit semantics, many works focus on using ontology to use in disambiguation process. [Angeletou 2008] and [Cantador 2008] associated tags with ontology. [Szomszor 2008] proposed an approach by assigning Wikipedia URIs to tags to disambiguate. Passant et al. in [Passant 2008a], introduced MOAT (i.e. Meaning Of A Tag) a lightweight Semantic Web framework that provides a collaborative way to let Web 2.0 content producers give meanings to their tags in a machine readable way. They construct an ontology in order to give for each tag its possible meanings and each meaning is related to a Linked Data concept from DBpedia. [Gracia 2009] gives an overview of a multi-ontology disambiguation method, aimed to discover the intended meaning of words in unstructured web contexts. The method takes an ambiguous keyword and its context words as input and provides a list of possible senses for the keyword, each meaning is given a score according to the probability of being the intended one. It accesses the online ontologies as source of word senses (e.g. WorldNet [Miller 1995]) to process its disambiguation

algorithm. [Khelif 2008] presented an ontology-driven word sense disambiguation process. The main idea is using the context of the ambiguous word to choose which class of the ontology to be assigned to. The disambiguation process relies on similarities between classes assigned to the ambiguous word, classes assigned to terms close to it in the text, and on the type of properties that could occur between them. However, this approach requires a text long enough to retrieve this kind of properties.

- Context based approach Other works focus more on the aspect of context, for example, in [man Au Yeung 2007], authors attempted to develop effective methods to disambiguate tags by studying the tripartite structure of folksonomies (i.e. namely users, tags and resources). [Lee 2009] proposed a disambiguation method, called Tag Sense Disambiguation (TSD), TSD can be applied to the vocabulary of social tags, thereby enabling users to understand the meaning of each tag through Wikipedia. In order to do that, they define the Local Neighbor tags, the Global Neighbor tags, and finally the Neighbor tags that can serve as keywords for disambiguating the sense of each tag. Another way to have context is to look at user's interest and intention. [Noor 2009] proposed a system that gathers information about users from their social web identities and enriches it with ontological knowledge. An interest model for the user can serve as a good source of contextual knowledge. This system extracts tags from social websites, filter them with WordNet and Wikipedia then mapping the concepts to general ontologies like YAGO[11] and DBpedia in order to map to more domain specific ontologies which support domain specific recommendations (e.g. Conceptual Reference Model (CRM) [Crofts 2006]).

  Another kind of approach in semantic disambiguation is presented in [Garcia 2009]. The authors proposed a context-based tag disambiguation algorithm that selects the meaning of a tag among a set of candidate DBpedia entries, using Cosine similarity - a common information retrieval similarity measure. The most similar DBpedia entry is selected as the one representing the meaning of the tag. The tags and the scores computed by the similarity are put in a repository called TAGora sense repository (TSR)[12] in order to be used as a dictionary where tags are related to DBpedia concepts and Wikipedia pages.

Sentiment Analysis

The case of sentiment analysis is somewhat similar, meaning that techniques can also be divided either in lexical or statistical with almost the same advantages and disadvantages as for entity extraction. An interesting resource for sentiment analysis is the

[11] YAGO - `www.mpi-inf.mpg.de/yago-naga/yago` - visited April 2011
[12] TSR - `tagora.ecs.soton.ac.uk/tag` - visited April 2011

SentiWordNet vocabulary, which associate to WordNet synsets the corresponding sentiment polarity. This vocabulary was built using a semi-supervised method [Esuli 2006a]. The main difficulty in using such a dictionary is the fact that a given term may appear in different synsets. Therefore, the right meaning must be selected before retrieving the sentiment polarity of the word. The most important problem in the case of entity extraction and sentiment analysis is the fact that these techniques were mostly designed for documents that have a well-defined structure. However, as mentioned before, social awareness streams are short with very little structure in terms of syntax. Therefore a corresponding research challenge is to establish how these methods must be modified or completely redesigned to work well also in the case of such micro-documents.

3.2 Semantic Metadata Management in Social Platforms

In the previous section we have explored existing work in the area of information extraction and processing of social content in Social Platforms (unstructured and structured). In the following, we consider the question of modeling, storing and publishing this extracted information. For this reason we explore the field of Semantic Metadata Management, that provides different models for the representation of connections between users and social objects.

3.2.1 The Semantic Web

Semantic Metadata Management was introduced by the the Semantic Web Community [Berners-Lee 2001]. The following quotation is considered to the be at its origin:

> "I have a dream for the Web (in which computers) become capable of analyzing all the data on the Web: the content, links, and transactions between people and computers. A "Semantic Web", which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize."

Although this vision targets more autonomous multi-agent systems and ubiquitous computing, it is however clear that its main objective is to provide web resources a unified description of their content for two main reasons: (i) interoperability, i.e. seamless exchange of data between systems and (ii) to offer web agents the capacity to interpret and reason on content. In order to achieve this, several technologies have been proposed and standardized over recent years. The most adopted are probably the Resource Description Framework (RDF) [Brickley 2004], a variety of data interchange formats (e.g. RDF/XML, N3, Turtle, N-Triples), and notations such as RDF Schema (RDFS)

[McBride 2004] and the Web Ontology Language (OWL) [The W3C Consortion 2010], all of which are intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain. An important reference of semantic metadata management specifically in the case of social networks is [Mika 2007b], where the author (P. Mika) investigates the mutual benefit of social networks and semantic web and more specifically, how the two can be combined together.

Currently, such technologies are not yet incorporated in Social Platforms on a large scale, as the adoption of the underlying practices is not easy for users (e.g. the use of common vocabularies for content sharing). Another problem is the fact that generally users have different viewpoints related to a given concept or domain, depending on their level of expertise, which makes it even more difficult to deploy a common vocabulary in a Social Platform for content management.

3.2.2 The Web of Data

Nowadays, more and more data is available on the Web. Such data can include information about organizations, governments, public people and from many other fields. More and more individuals and public/private organizations contribute to this deluge by choosing to share their data with others (e.g. Web-native companies, such as Amazon, Yahoo!, Flickr). Third parties consume this data to build new business opportunities and enhance recommendation strategies, such as online marketing and user re-targeting. Clearly, the further evolution of this ecosystem requires the presence of structured data, which certainly raises the question of how to provide access to data so that it can be most easily reused. The answer provided by the Semantic Web Community is based on the important principle that data published in the Web should take into account the fact that more structured it is, more easily it can be reused for different objectives and by different organizations and individuals.

The previously mentioned initiative, "The Web of Data" or "Linked Data", uses standard mechanisms for publishing and specifying the meaning of connections between the published items. More specifically, data is published using the Resource Description Framework (RDF), which was considered as the most flexible way to describe things in the world and specify meaningful relations between them. RDF allows to obtain with data the same as hyperlinks did with documents: link data islands stored on different servers into a global data repository. Beyond the use of this standard representation format, the following principles govern Linked data: (i) the use of URIs as names for things, (ii) the use of HTTP URIs so that these names can be accessible for people, (iii) the use of standardized access and data representations (RDF, Sparql) and finally, (iv) the inclusion of links to other URIs for efficient data discovery. The most important principle is probably the last one, as this is the key of the success of Linked Data. More concretely, such RDF links pointing to external data sources allows the Linked Data

graph to be interconnected.

Currently, the most complete dataset in the frame of linked data is certainly DBPedia [13], with more than 3.4 million concepts and relations extracted from Wikipedia. This dataset is in the category of cross-domain data (i.e. does not focus on a specific topic, such as music, geography or health), given the fact that things that are the subject of a Wikipedia article are automatically assigned a DBPedia URI. Therefore it has a very large topic-coverage and serves as a hub in Linked Data, as almost every item in other datasets are assigned an external link to DBPedia 3.5. Further cross-domain data sets include that of Freebase, Umber, Yago and OpenCyc. Almost each of them is linked to DBPedia. Other datasets target a specific domain to publish data. This includes geographic data (information about places, countries and locations) with GeoNames, containing data about almost 8 million locations. Almost each location item in Geonames is linked to the corresponding concept in DBPedia, which allows to retrieve additional information about the place (such as nearby locations or people related to the place or its category). Fields that are also well covered by such datasets include government data (e.g. data.gov.uk, data.gov), libraries and more generally, education (e.g. OpenLibrary, Semantic Web Dogfood Server). Less significant datasets are also available for life sciences, such as biology (e.g.Bio2RDF) and commerce (e.g. RDF Book Mashup). All of these datasets and the wealth of incoming and outgoing links between them clearly show the diversity of Linked Data and its potential for advanced recommendation strategies and web resource discovery.

In the following chapter we will further depict the internal structure of Linked Data. For this reason, we will specifically focus on the different Semantic Web vocabularies that can be employed to publish information in Linked Data and also how these vocabularies evolved and their targeted scope.

3.2.3 Semantic Metadata Management - Capturing User-Centered Information in Social Platforms

Given the huge amount of shared content, metadata plays increasingly a central role in Social Platforms. Generally speaking and in the context of Social Web, metadata is any kind of data that provides additional information about a resource or an annotation. Their effective exploitation requires methods and tools that facilitate their management. This includes traditional issues such as modeling, specification, generation, curation, storage and retrieval, that are typical for any data management system. Over recent years, the Semantic Web had an increasing contribution in metadata management. This resulted in the term semantic metadata, which consists of metadata that is defined using shared conceptualizations (e.g. ontologies) or other kinds of controlled vocabularies (e.g. taxonomies or object models). The advantage of semantic metadata is clear, as

[13] DBPedia, `www.dbpedia.org` - visited January 2011

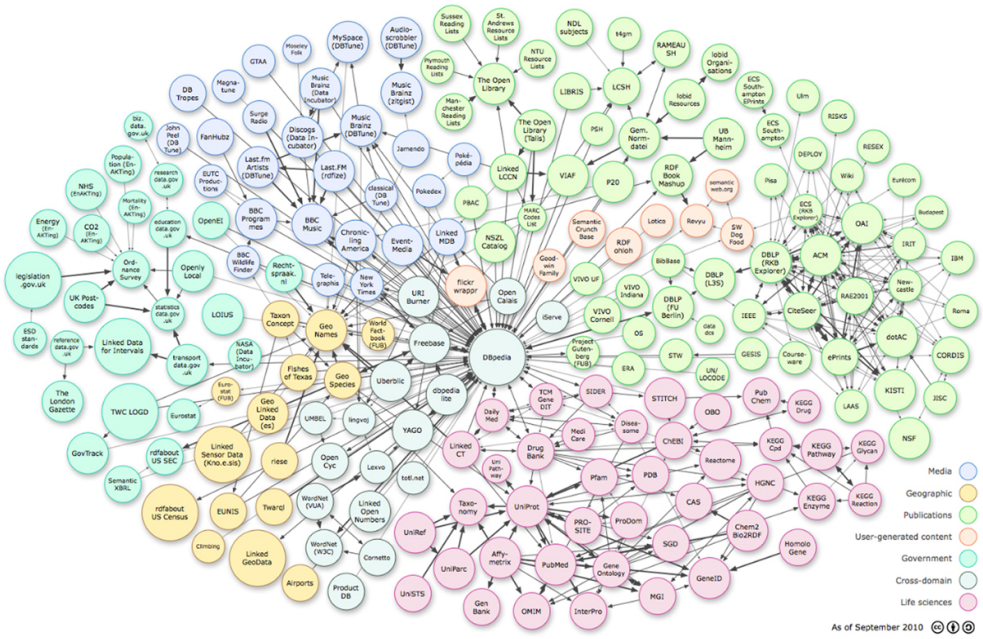


Figure 3.5: Linked Data Cloud (Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch `http://lod-cloud.net/`)

a shared conceptual definition is a key for interoperability of social content between Social Platforms.

Currently, from the metadata and content management perspectives, social platforms generally have proprietary models for the storage of information produced by users (e.g. profiles and shared content). This limits the reuse of the social content between the different platforms, a phenomenon usually conceptualized with the walled garden metaphor, meaning that such systems are separated and data is not exchangeable. The rational behind the necessity of going beyond walled gardens in terms of social content management is the fact that users are often members of multiple social platforms. Therefore a flexible interchange of information between such systems would significantly facilitate the management of identities and shared content (e.g. seamless reuse of user profile information, of tags on photos or status updates etc.).

The Database and Semantic Web Communities are particularly active in proposing solutions to overcome this problem [Shadbolt 2006] [Lakshmanan 1996]. In this chapter we focus only on the contributions of the Semantic Web Community, which proposed several models to establish an open and standardized representation and storage of social content. It is also interesting to see how social content is processed from the perspective of these technologies. However, with the increasing number of proposed models, it is very hard to select a specific model without being constrained to review most of the existing ones. In fact, few information is present in the literature about the contribution of each model and its specific application area and no comparison is available that clearly states the differences and relations between them.

The Semantic Web community introduced several models for metadata management. Such models target either individually a given component of the system, such as the user and her profile, the social objects or the combination of both. In the latter case, we speak about general social web models. In the next section we review the most important models in order to understand the main concepts they user for modeling the main components of online communities and to understand the specificities of each model (Figure 3.6).

Our review of semantic metamodels completes previous works in this field done by [Garcia-Silva 2011], which focuses on the review of models that deal with the association of semantics to tag data and work by [Kim 2008b], which reviews these models from a structural perspective (i.e. structure of composing ontologies).

3.2.4 Semantic Models for Capturing Social Content

After a brief introduction of the different pillars and their related components, this section focuses on the models proposed by the Semantic Web community. We first focus on models covering the capture of information about users (User Profile Ontologies), shared content (Tag Ontologies) and then more general social web models. Secondly,

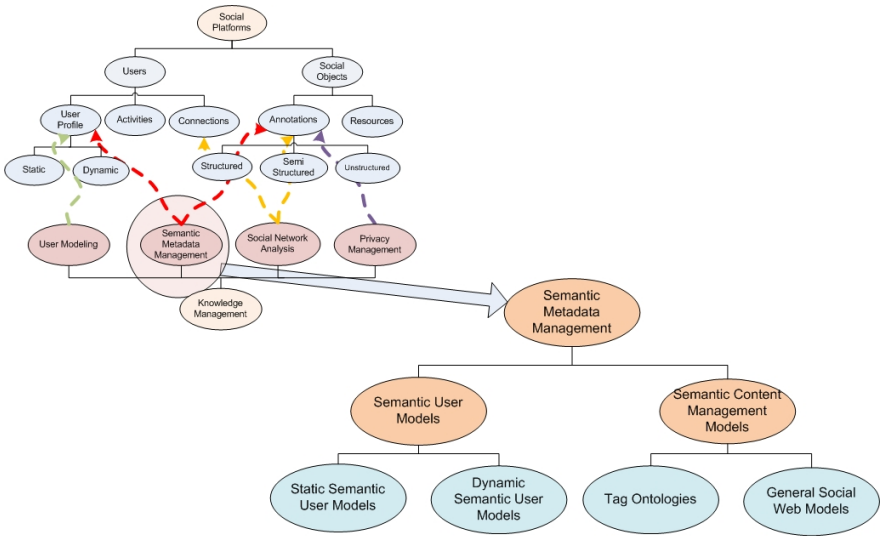


Figure 3.6: Main dimensions of Semantic Metadata Management

we compare the proposed models using a set of criteria specially tailored to facilitate the selection of the best model to a specific need in software design.

3.2.5 User Profile Ontologies

Most of the early user models are mainly related to profile representation [Gauch 2002] [Trajkova 2004] [Teevan 2005][Katifori 2007a]. In other words, the objective is to model the information related to the user in order to build a realistic picture about his/her context and interests while making it portable from one platform to another.

In his book, J. Windley [Windley 2005] mentions three identity tiers in the case of components for a user profile model:

- *My identity* (persistent information that is static, timeless and unconditional, i.e. name, color of eyes).
- *Shared identity* (attributes assigned to a user by others, i.e. social network information, such as "expert in semantic web").
- *Abstract identity* (information derived from groupings or temporary labels, i.e. status information, current location).

In the following we show how semantic models consider these identity tiers. The most known semantic model for user profiles is certainly *Friend-of-a-Friend* (FOAF). FOAF is an RDF vocabulary targeted to describe profiles, friends, affiliations, creations,

work history, links to contacts and services, etc. FOAF intends to provide a way to representing information about people in a way that is easily processed, merged and aggregated. The spirit behind FOAF is that of a completely decentralized machine-readable social network that is based on personal profiles. From a usage perspective, a person begins by describing herself using the $foaf : Person$ class, listing key attributes such as name, gender and related resources. They can also list their interests, and each person is uniquely identified by using the $foaf : mbox$ property containing her email address[14] The person moves then to describing her friends. Each friend is an instance of the $foaf : Person$ class. FOAF is going through a slow adoption by Web 2.0 sites and services, this could be explained to the lack of interest in the exportation of social information from one rival site to another and thus going beyond the walled gardens paradigm. It is, however, the most widely used specification for expressing personal and relationship information within the Semantic Web community. Existing systems that adopted FOAF on a large scale are the following: the Hi5 [15], the Mybloglog community builder service of Yahoo [16], FriendFeed [17] and the Typepad blogging service [18]. In all cases, an API allows to export social data in the FOAF vocabulary. Other systems using FOAF include *identic.ca* [19] and *Drupal7* [20].

Microformat (sometimes abbreviated μF or uF) is a Web-based data representation vocabulary that allows reusing existing content as metadata, using only XHTML and HTML classes and attributes. This approach is intended to allow information for end-users, such as contact information, geographic coordinates, calendar events, etc. to also be automatically processed by external software. Microformats allow the encoding and extraction of events, contact information, social relationships and so on. Most important uses of this technology is the integration of semantic contact data in Web pages. Such examples include vCard (Electronic Business Card) [Consortium 1995] and hCard (vocabulary for representing people, companies, organizations, and places [John 2007].

A number of compact formats, variously called nanoformats [21], picoformats [22], or microsyntax [23] have been proposed to allow users to express structured content or issue service specific commands in microblog posts. Examples in widespread use include *@usernames* for addressing or mentioning a particular user, and #hashtags for generic concepts. So-called triple tags even allow the expression of something like an RDF

[14] An alternative identification property is $foaf : openid$ conforming to the *OpenID* [Fitzpatrick 2005] initiative of using a unique single URI to establish the identity.

[15] Hi5 Social Networking Platform, www.Hi5.com visited November 2010

[16] Community Builder, www.mybloglog.com - visited November 2010

[17] Social Awareness Stream Aggregator, www.Friendfeed.com - visited November 2010

[18] Blogging Service, www.Typepad.com - visited November 2010

[19] Social microblogging Service, www.Identi.ca - visited November 2010

[20] Content Management System, www.Drupal.com - visited November 2010

[21] http://microformats.org/wiki/ microblogging-nanoformats

[22] http://microformats.org/wiki/picoformats

[23] http://www.microsyntax.org/

triple. These formats are subject to a tradeoff between simplicity and expressivity which heavily impacts community uptake.

Carmagnola et al. [Carmagnola 2007] investigate how tagging helps to infer data about user preferences or interests, allowing to create a user model. Tagging is the process where users label or annotate different resources (Web-pages for example) with the objective to share, organize or diffuse them. This is a concept strongly related to the Web 2.0 and social networking. The way users employ tags might give an insight on different issues like how interested they are in the given resource and the type of tags used (many synonyms for example) can infer subjective details like level of creativity. Tests and deep analysis are needed to better understand relations between tagging and these high-level concepts (preference, interest). Research is currently in progress examining how ontologies like Wordnet allow categorizing and automatically inferring the type of tag and relationship with the user (matching).

Von Hessling et al. [von Hessling 2004] propose an architecture where semantic user profiles are used in a peer-to-peer mobile environment. In a ubiquitous mobile environment, services providers like a cinema are equipped with bluetooth-enabled devices to broadcast the service (movie being played for example). People have mobile devices which store the owner's profile (interests, preferences, disinterests). Easy matching between user interests and services is necessary, and description logic is considered a good approach. The user profile is relatively simple, consisting simply of the union of interests and disinterests: common domain ontology for concepts in both services and profile description is used to make possible this semantic matching using a "reasoner". The system is completely peer-to-peer, allowing independence, cost effectiveness, scalability and most importantly, privacy management, which is the result of the fact that profiles are stored on the mobile device.

Rousseau et al. [Rousseau 2004] argue that techniques like RDF and OWL together with ontologies are the key elements in the development of the next generation user profiles. The User Profile Ontology presented grew out from a quite simple model containing semantic contact information encoded in the RDF language. In the proposed ontology which is built around three imported ontologies (*Person Ontology*: containing classes relevant only to the user; *Organization Ontology*: containing business oriented information, and *Common Ontology*: containing information relevant to both persons and organizations) personal information like e-mail, phone numbers, Instant Messaging identifier, physical addresses are uniquely identified by a $GenericContactIdentifier$ class. Social interactions inherit properties from an *Event* class. These interactions are classified into voice, text, and real-time, online communications. The address book of the user is stored in the class *ContactGroup*. The applicability of the User Profile Ontology is tested with a Dynamic Address Book application, which maintains dynamically the contact list of a user based on the frequency of communications. Rules are defined to decide when to remove or add a contact according to a given calling frequency

threshold. Results show that this is a promising approach, but further evaluation of the system is needed.

A second category of work focuses more on capturing dynamic information about users that relate more to the shared and abstract identities. Katifori et al. [Katifori 2007b] present an application-independent user profile ontology. The objective is to create a "general, comprehensive and extensible" user model taking into account user- and context models existing in the literature. The use of ontologies in user profiling is a known issue, however the problem with existing proposals is that they respond only to application specific needs, mainly in personalized information retrieval and Web search. It is important to stress out that the proposed ontology deals only with the static profile of the user, not the dynamic or contextual one (like current position, occupation or terminal). The proposed ontology, where the main class is Person, uses these concepts and many others to create a profile applicable in any kind of domain or application. Therefore, no restrictions are present in the ontology; it is completely up to the developer to personalize it according to the specific needs of the project. However, some restrictions could have been added like the fact that a friend of a Person can only be a Friend, which is still domain independent and represents general knowledge.

Vildjiounaite et al. [Vildjiounaite 2007] address the issue of modeling users in a context-aware "smart home" environment in the context of the *Amigo* project. The user model is separated into two components: (i) the context-aware static user profile (Tree-based representation of individual user preferences and personal data, grouped in agreement with user ontology representation in the system), and (ii) The context-aware dynamic user profile. This profile, i.e., context-aware dynamic user profile, learns user behavior from history of activities, learning meaning the ability to recommend a given topic in a given situation (for example a movie when Bob is alone at home on Friday night). Interaction history is stored in the form *"Context"→"User Action"*, where context is a set of environment descriptors and user action any kind of interaction with the "smart home". Two machine learning techniques, CBR (Case-based reasoning) and SVM (Support Vector Machines) were used to test how items can be classified into terms like "good" or "bad" for a given context. Test results indicate that CBR gives good performances in almost any situation, since the retrieval of items is based on the overall similarity of context descriptors.

A recent model that focuses on the abstract identity category is the *Online Presence Ontology* (OPO) [Stankovic 2008]. The main novelty in this model is the inclusion of geo-location information. This is based on the observation that when people are on the move, permanent locations specified in their user profiles do not help much to position them when sharing content. OPO allows modeling the current activity of the user and specifying reachability restrictions.

Finally, another interesting work in this category is the *User-Profile Ontology with Situation-Dependent Preferences Support* (UPOS) [Sutterer 2008]. This model ad-

dresses both static and dynamic aspects. This ontology, defined in OWL, allows creating situation-dependent sub-profiles. A user has a profile and a context (location or activity) associated. The notion of condition is defined, which includes a user, an operator and a context-value. For example, a condition can be: "*if the context of user Bob equals the MyOffice location*". According to this condition, a corresponding sub profile can be applied that contains all personalization indications for services (e.g., not to use SMS). The most important guideline of this approach is to structure the profile into sub-profiles, each containing user preferences that correspond to a specific situation, as seen in the previous example.

3.2.6 Semantic Models for Capturing Shared Content

Content is certainly the part which has attracted most attention in the area of the social web. Currently numerous semantic social metadata models exist (*Gruber, Newman, Knerr, MOAT, SCOT, CommonTag, NiceTag and SIOC*). The early models from Gruber [Gruber 2005] and Newman *Tag Ontology* are at the basis of most semantic content management models. These models are not integrated in web applications, but are still useful as they allow to understand the main ideas behind semantic content management and its main dimensions. Two main categories of models exist in this category: (i) Tag Ontologies and (ii) General Social Web content management models.

3.2.6.1 Tag Ontologies.

A first effort to conceptualize the activity of tagging resources on the Web is that of Gruber [Gruber 2005]. The model is not an ontology but rather a general representation model for tagging. Gruber's Tag Ontology attempts to create a common representation and an infrastructure to link taggers, tags and resources. Its objective is not to provide a common vocabulary of keywords to use for tagging. More concretely, the core concept of this ontology is an abstract term of the tagging activity itself, called "Tagging". This is a formalization of a social activity, i.e. that of labeling some content with one or more tags, e.g. freely selected keywords. By considering also the collaborative aspect of Web. 2.0, i.e. more users can tag the same resource. This Tagging activity is an association between an object, a tag and a tagger: $Tagging(object, tag, tagger)$. By defining the source as the scope of name-spaces or universe of quantification for objects, the model allows one to differentiate between tagging data from different systems and is the basis for collaborative tagging across multiple applications. Mika [Mika 2007a] already represented the tagging action from a theoretical point of view, but did not use this notion of source that Gruber introduces. Yet, while this model is widely commented, there is no currently available implementation to the best of our knowledge.

Newman et al. [Newman 2005] defined an ontology of tags and tagging, simply called the *Tag Ontology*, that describes the relationship between an agent (tagger), an arbitrary

resource, and one or more tags. Thus, in this ontology, the three core concepts Taggers, Tagging, and Tags are used to represent the tagging activity. Contrary to Gruber, it does not represent the source of the tagging action. Notably, in this ontology tags are represented as instances of the tags:Tag class which is assigned custom labels, i.e. the string representing the tag as seen by the user. Being instances of a class means that they are assigned an URI. URIs are key features of the Semantic Web, since, contrary to simple literals, they can be used as subject of triples, while literals can be only used as objects. This way, tags-identified by URIs can be linked together and people can semantically represent connections and similarities between tags. For this purpose the ontology introduces a $tags : related$ property. This ontology has been implemented (in OWL) and is available on the Web. It is currently used in some projects such as *Revyu*, a review system combining Web 2.0 and Semantic Web technologies.

Tagging Ontology is another interesting approach proposed by Knerr [Knerr 2006]. In the proposed ontology, the central element is the concept of Tagging, which is considered as being a tuple of *(time, user, domain, visibility, tag, resource, type)*. The particularity of this model is its visibility concept. The objective of assigning to a tag a visibility option is to allow for public, private and protected tagging. Public tags are visible for everyone, private tags only for the tagger himself and protected tags are meant to be visible for a selected group of people (e.g. friends). Some other features of this ontology include the consideration of several mappings with Gruber's model concerning the identity of tags and with Newman's model with the use of the SKOS vocabulary for tags (described hereafter).

Meaning of a Tag (MOAT) [Passant 2008a] attempts to provide a Semantic Web model to define the meaning of tags in a machine-readable way. To achieve it, MOAT defines: (i) the global meanings of a tag, i.e. the list of all meanings (ii) the local meaning of a tag, i.e. the meaning of a tag in a particular tagging action. For instance, the tag "Paris" can mean - depending on the user, context and other factors - a city in France, a city in the USA, or even a person. Yet when someone uses it in a tagging action, it has a particular meaning, for example the French capital. Thus, MOAT extends the usual tripartite model of tagging action to the following quadripartite model Tagging (User, Resource, Tag, Meaning). Moreover, MOAT introduces a social aspect that lets people share their tags, and their meanings, within a community by subscribing to a MOAT server, as they could do with the Annotea annotation server [Kahan 2001]. They can share and update tag meanings, and use it when tagging content. When a user tags content, the client queries the server to retrieve tag meanings and let the user choose which one is the most relevant one, regarding the context.

The Social Semantic Cloud of Tags (SCOT) [Kim 2008a] aims to describe the structure and the semantics of tagging data and to offer social interoperability of the data among heterogeneous sources. According to the authors, the main design principles of SCOT are the following: (i) lightweight data representation (e.g. RDF),

(ii) enhanced capabilities for sharing tags among users, and (iii) compatibility with existing lightweight vocabularies.

Simple Knowledge Organization System (SKOS) [Miles 2008] is a RDF vocabulary for representing semi-formal Knowledge Organization Systems (KOSs), such as thesauri, taxonomies, classification schemes and subject heading lists. Because SKOS is based on the Resource Description Framework (RDF) these representations are machine-readable and can be exchanged between software applications and published on the Web. SKOS has been designed to provide a low-cost migration path for porting existing organization systems to the Semantic Web. SKOS also provides a lightweight, intuitive conceptual modeling language for developing and sharing new KOSs. It can be used on its own, or in combination with more-formal languages such as the Web Ontology Language (OWL). SKOS can also be seen as a bridging technology, providing the missing link between the rigorous logical formalism of ontology languages such as OWL and the chaotic, informal and weakly-structured world of Web-based collaboration tools, as exemplified by social tagging applications.

CommonTag Ontology [AdaptiveBlue 2009] is an open tagging format developed to make content more connected, discoverable and engaging. Unlike free-text tags, Common Tags are references to unique, well-defined, and complete concepts with metadata and their own URLs. With Common Tag, site owners can more easily create topic hubs, cross-promote their content, and enrich their pages with free data, images and widgets. The Common Tag format is based on RDFa [24]. The Common Tag structural model is very simple. It states that a piece of content addressable through a URL (a "resource") can be tagged with one or more Tag structures. Each tag can contain a pointer to another resource that identifies the concepts described by the content, unambiguously indicating what the content means. Optionally, the Tag may also contain information about when the Tag was created (the tagging date) and what human readable label should be used when listing the concepts covered in the content.

NiceTag Ontology is a recent model that describes tags with RDF named graphs [Limpens 2009]. The main observation for the necessity to introduce named graphs in tag modeling is the fact that current approaches do not take into account that tags can take different forms and be of different nature. Thus, a model is needed that allows describing tags in a very general way. The NiceTag Ontology is particularly rich as it allows expressing different intentions behind a tagging action and contains basic privacy manipulation concepts. More concretely, a user can define the fact that a given tag expresses a feeling (*ExpressFeelings*), points to part of a Web resource (*Point*), is used to limit access to a resource (*GiveAccessRights*) or to evaluate it (*Evaluate*). Each tagging action is linked to the user who performed it. For this, the ontology

[24] RDFa is a standard mechanism for placing structured content within HTML documents. The RDFa standard was developed by the W3C and is supported by both Yahoo! and Google as a way of providing additional information about the page to search engines (http://www.w3.org/TR/rdfa-syntax/)

makes the distinction between the owner of the resource (*OwnerTagAction*) and other users who visit the Web resource and tag it (*VisitorTagAction*). Other main features of this ontology include concepts allowing to specify that a tag is disambiguated, used to share a resource or for indicating that a future action is required with the resource. Also, the ontology allows to deal with collections of tags, either on individual (*PersonalTagCollection*) or community level (*CommunityTagCollection*).

Semantically-Interlinked Online Communities: SIOC [Breslin 2005] is a vocabulary that aims at interconnecting on-line community sites and Internet-based discussions. Although not a Tag Ontology, this model is still important as it gives a more general model of social Web content management. The idea is to enable cross-platform interoperability so that conversation spanning over multiple on-line media (e.g., blogs, forums, mailing lists, etc.) can be unified into one open format. The interchange format expresses the information contained both explicitly and implicitly in Internet discussion methods, in a machine-readable manner. On-line communities allow Web users to express their thoughts, gain feedback and interact with individuals who share a similar interest. Modern Web users all have some kind of participation in this realm: forums, chat rooms, newsgroups and social networking sites. Each community can be considered as a walled garden, without link to others. The SIOC project focuses on ways to integrate and merge these walled gardens, providing bridges between the users and resources that exists in each of them.

Other proposals for capturing social content include the following: LODR (A Linking Open Data Tagging System) [Passant 2008b], which is very similar to MOAT, the GroupMe! ontology from Abel et al. [Abel 2008a], that adds additional contextual information to the tagging activity (e.g. time, mood of the user, background knowledge of the user etc.) and the model designed by Echarte et al. [Echarte 2007], which considers an additional dimension, the sentiment polarity of tags.

3.2.7 Semantic Resource Management

A resource, as explained in the previous sections, can represent any object of interest such as photos, videos, documents, discussions, etc. One of the most interesting properties of the Semantic Web is to reference resources, users and tags with URIs (i.e. Uniform Resources Identifier). The concept itself was first introduced by Berners-Lee. URIs are identifiers of information objects on the Web, by giving them an address where users or Web crawlers can access them. Objects on the Web that have an URI are called "First Class Objects". The main properties and principles are the following: (i) Universality (i.e. any resource anywhere can be given a URI and any resource of significance should be given a URI),(ii) Global scope (i.e. it doesn't matter to whom or where you specify that URI), (iii) sameness (i.e. an URI always refers to the same thing), and (iv) Identity (i.e. the significance of identity for a given URI is determined

by the person who owns the URI, who first determined what it points to).

3.3 Analysis of Semantic Metadata Management Models

As seen in the previous section, the Semantic Web community designed a series of metadata management models, each covering specific aspects of social content. In this section we present our analysis dimensions and a comparative table according to these criteria. The next section will detail each model and explain the different scores affected.

Table 3.1 provides a summary of the analysis where columns represent the different considered models and rows the considered analysis criteria (i.e., dimensions).

The following notations have been used in the table: (i) columns represent the semantic user and content management models, the comparison criteria, (ii) each cell of the table contains a rating: "+" sign means that the given model provides/supports a considered criteria, "×" means that the considered criteria is the (one of the) most prevalent characteristic of the corresponding model and can eventually be considered as a reference for the given criteria. On the contrary, the "-" means that the given model does not deal with the considered criteria at all. This should not be considered automatically and uniquely as a lack of a particular model but rather as simply the non consideration aspect.

The analysis is performed with respect to three focused categories of criteria and an additional general category. A first category of criteria attempts to measure how well a given model allows to inter- or intra-connect these components either (social connectivity, resource connectivity, tag connectivity). As annotations and, more concretely, social tags are the most widely used annotations in the Web 2.0, a second category of criteria attempts to capture the capacity of a model to manipulate these tags, i.e. tags manipulation. The third category of criteria provides insights of how well a given model addresses privacy and security issues. Finally, a general category is considered which includes metadata about a given model (format, extensibility, current usage). In the following we describe more concretely the different analysis axis and discuss the different presented models with respect to each axis.

3.3.1 Connectivity of the model

Translated by the *Activity* axis, this dimension intends to capture the strategy of a given model to link specific entities of the same nature in the eco-system or even interlink them, i.e., users, tags, and resources.

For example, the *social connectivity* relates to the capacity of the model to build connections between users. More specifically, it examines whether the model allows for users to create explicit or implicit communities (i.e. groups of users), whether a user can belong to one or more communities and how these communities can be connected

Table 3.1: Summary of the considered dimensions and the affected scores for each model regarding the analysis dimensions

		FOAF	**OPO**	**UPOS**	**Knerr**	**MOAT**	**SCOT**	**SKOS**	**CommonTag**	**NiceTag**	**SIOC**
Connectivity	*User*	×	-	-	-	-	+	-	+	+	+
	Resource	-	-	-	-	+	+	-	+	×	+
	Tag	-	-	-	-	+	×	+	+	+	×
Tag Semantics		-	-	-	-	+	-	×	+	+	-
Tag Aggregation		-	-	-	-	-	×	+	+	-	+
Privacy Management		-	×	-	×	-	-	-	-	+	-
Tagging Rights	*Self*	-	-	-	-	+	+	+	-	+	-
	Free-for-all	-	-	-	-	+	+	+	-	+	+
Vocabulary	*OWL*	+	+	+	+	+	+	+	+	+	+
	RDF	+	+	-	+	+	+	+	+	+	+
Portability		×	+	+	+	+	+	+	+	+	+
Usage Frequency		×	-	-	-	-	-	-	×	+	×
Extensibility		×	×	+	+	+	+	+	×	+	×

between them. Also, an important consideration is whether the model allows only connections between users or also indirect connections, e.g. resource-centered. More concretely, two users can also be connected, by e.g. the fact that they annotated the same resource.

Similarly, *resource connectivity* considers whether or not the resources in the system based on a given model are linked to each other independently of user's tags, and how these resources are connected. In other words, this dimension expresses if collections of resources can be established using a given model.

Finally, *tag connectivity* relates to whether or not the tags/annotations in the system based on a given model are linked to each other independently of user's tags, and how these tags are connected. More concretely, we will examine in this case if collections of tags (e.g. with similar semantics) can be established using a given model.

It comes from the comparison that most of the proposed models aim mainly at linking specific entities of the ecosystem between them. *NiceTag* is the most interesting model from this point of view since it enables linking both users and resources. *FOAF* focuses on the capture of relations between users which is the main objective behind it. Finally, *SCOT* is the only model which is strongly operating on tags linkage. The fact that most of the models don't consider necessary the connectivity is due to their objective which is generally descriptive rather than networking.

3.3.2 Social tag description

This is mainly translated by the two axis: *tags semantics* and *tags aggregation* which are respectively: (i) the management of the meaning of tags were considered in the system (e.g. the possibility to link the tag to an ontology that describes its meaning or the context in which the tag was used to annotate a resource or user). (ii) the allowance of a multiplicity of tags for a given resource, which is called *bag model*. *The set model* is the case where a group of users can collectively tag a resource (this avoids the repetition of tags). As content-based recommendation represents nowadays the major focus of research in the Social Web, this is a very important criterion to classify and rank metadata management models. Models that explicitly deal with this issue are *SCOT*, *SKOS*, *MOAT* and *CommonTag*. *SCOT* has a particular way of modeling the tagging activity, as it considers a Tag Cloud (concept *TagCloud*) as the core concept. The *SIOC* vocabulary (concept *UserGroup*) is used to define the group of users who performed the tagging of a given resource. In this group, each user can be described with the *User* class, which points to the *FOAF* vocabulary for a detailed description of the user's identity (static profile). An individual tag is associated to a Tag Cloud by a tagging activity concept (*TaggingActivity*). For each Tag in a tagging activity, a number of properties allow to specify the author and some collective features, such as the frequency of a tag in the given tag cloud and statistical measures of co-occurrences. Also,

there exist properties for tag aggregation. Although concepts for expressing semantics are not directly included in the model, we consider that it is easy to connect other vocabularies, e.g. *SKOS* for this purpose. This is a model to use if a given application requires the consideration of statistical information about social data, like co-occurrence measures or aggregations.

In the case of *SKOS*, its main power is the fact that it allows to express semantic relations between tags, allowing their organization into formal hierarchical structures (e.g. taxonomies). This is achieved thanks to a number of properties, like *broader, narrower*, and *related*. The first two allow expressing a kind of semantic distance between two tags (hypernymy and hyponymy relations) where the last allows connecting tags with similar meaning (synonyms). Based on these observations, this vocabulary is very interesting for, e.g. recommendation strategies that attempt to leverage the meaning of social data. This vocabulary can be connected to, e.g. *SCOT*, to have a complete federated metadata management model, that includes structural and semantic information about tags in the same time. For all this reasons, *SKOS* is considered the strongest vocabulary in terms of semantics.

Finally, *MOAT* is also a vocabulary that considers semantics. It does it in a different way, by providing tags' disambiguation with the association of a URI that points to a concept from a knowledge base (e.g. DBPedia). This is achieved by the property *moat:Meaning*. *CommonTag* offers the same feature with the property *ctag:means*. This is very important in the case of tags that can have multiple meanings (e.g. apple - company or fruit). As disambiguation is key for relevant recommendations to the user, MOAT and CommonTag are considered the second in our ranking related to semantics. As in the previous case, *SKOS* can be connected to *MOAT* or *CommonTag* to allow this time a full conceptual framework for semantics, including both hierarchical and synoymy relations between tags (from *SKOS*) and disambiguation capacity (from *MOAT* or *CommonTag*).

3.3.3 Privacy and Security Management

This is translated by the two axis: *privacy management* and *tagging rights*. The privacy aspect deals with one of the most important issues in the Web: the privacy of users, tags and resources. More specifically, different issues related to privacy management need to be considered: do users have the possibility to specify the visibility of their tagging data (e.g. to whom the tags will be visible?). Regarding the tagging rights, A tagging system can be restricted to *self tagging*, where users tag resources they created. Another type allows *free-for-all* tagging, where any user can tag any resource. Between the two limits, there are several variants: the system can choose the resources users can tag, or specify different levels of permission to tag a resource. Similarly, the way tags are removed can have different variants: no one, the owner, the viewer, or different

social relations (friends, family, etc.). Also, tags assigned to a photo for example can take different forms according to whether the creator is the owner, a friend or a stranger. A first conclusion is that only one Tag Ontology deals with a very important issue on the Web 2.0: the privacy of the data. Therefore, if application developers intend to integrate privacy management issues (e.g. to restrict the visibility of a tag according to social categories), they should reuse and extend Knerr's model. Knerr proposes a very basic visibility restriction mechanism, but it is easily extensible and sufficiently portable, as most part of the model overlaps with Newman's Tag Ontology. For example, this ontology would be sufficient to provide a semantic presentation layer for tags together with the tagged resources and their visibility for Facebook. Using this ontology, users could export their tag clouds and import it to another application (e.g. Flickr or other social network).

Also, it is important to note that the proposed mechanism for privacy management is very basic, and does not deal with several dimensions described in the Data Privacy Taxonomy [Barker 2009]. More clearly, dimensions like granularity, also pointed out by [Golder 2005], retention are not considered in the model. Granularity refers to the possibility to share different levels of granularity of social content based on the social category of people who visualize the information. A connection of SKOS with SIOC communities would allow to specify such kind of granular sharing preferences. The Online Presence Ontology considers also this dimension, but in this case the concepts are related to the on line presence of the user and not the social data (e.g. "Findability" (instances: *PubliclyFindable, ConstrainedFindability*), "Notifiability" (instances: *AllNotificationsPass, NotificationsConstrained, NotificationsProhibited*). More concretely, these concepts allow a user to describe her on line presence and to manage interruptions. Finally, the NiceTag ontology includes some basic concepts and properties for privacy management of tags, such as the properties *cannotBeReadBy* and *canBeReadBy* that can be used to indicate to whom access right to the tagged resource are denied or allowed.

3.3.4 General criteria about vocabulary and usage

These dimensions consider the rest of the dimensions indicated in Table 3.2: (i) *portability* referring to the presentation layer of the proposed ontology and measures whether existing ontologies are included in the model. For example, if a concept, property already exists in ontology, it should not be recreated with a new name/name-space. (ii) *extensibility* measuring the capacity of the model to allow coupling other ontologies (e.g. context or user model ontologies). Coupling a context ontology (e.g. SOUPA, Cobra, Cool, Gaia, CONON [Baldauf 2007]) can be useful for example to better explain the context of the tagging activity (e.g. location about the user etc.). A user model can help in better describing the user or relations between users in the tagging system. (iii)

Usage of the model in the Social Web which examines whether a given model is already used in some Social Web applications and what is the feedback of users. Finally, (iv) *vocabulary* translating the representation formalism used in a specific ontology. We can notice the two main formalisms used are *RDF* and *OWL*. We can notice that most of the models support *RDF*.

The most important benefit of such tag ontologies is clearly application oriented and they provide an important component for reasoning mechanism. Users themselves rarely manipulate directly folksonomies. Reasoning mechanisms could use such semantic representation to create links between users, resources or tags and guide the user in the tagging process. Generally a semantic meta-model is designed to answer a specific requirement related to some user-related action on the Social Web (e.g. sharing a resource, disambiguating meanings of tags, managing the visibility of tags). A complete flow of requirements, cannot be achieved by one single model. Mashups (e.g. combinations or alignment) of different models are thus needed in order to cover all collaborative tagging - related requirements of a social application. This process is also called ontology alignment. The main difficult however is not the alignment itself, as all models are expressed in standardized and widely accepted formalism (e.g. RFD, OWL). The main difficulty resides in choosing the right model for a specific requirement, such as the management of the visibility of user generated content or the disambiguation of semantics in the case of the meaning of tags.

3.4 User Modeling in Social Platforms

After the review of contributions to knowledge management from the Social Network Analysis and Semantic Web fields, we focus our attention in the following on users. Users are indeed the main actors in virtual communities. They are composed of a profile, activities and connections. The modeling of users has been an important preoccupation of various research communities (e.g. User Modeling and Personalization (UMAP), Computer Human Interaction (CHI)). The increase of the variety and complexity of social platforms increases and therefore the understanding of how the system can dynamically capture relevant user needs and traits, and automatically adapt its behavior to this information, has become critical. A good user model is important for relevant recommendations and filtering.

Formally, any user model is represented by a set of concepts and associated weights. We consider U the domain of all users involved in the social platform. CU represents the set of items correlated with user u, i.e. $CU = c|P(c,u) > 0$. Therefore, user u and item c are correlated when $P(c,u) > 0$, P being the weight of the item in the profile (representing the degree of interest the user shows towards the given topic).

In the literature we distinguish two main types of user profiles [25] (Figure 3.7):

- keyword based [Carmagnola 2007] (i.e. the user profile is a bag of words, no effort is considered to associate meanings to the items).
- concept based [Vallet 2007] (i.e. the user profile is a kind of bipartite graph, as each item is associated to a concept from a knowledge base, that disambiguates its meaning. Thus, further relations may exist between concepts).

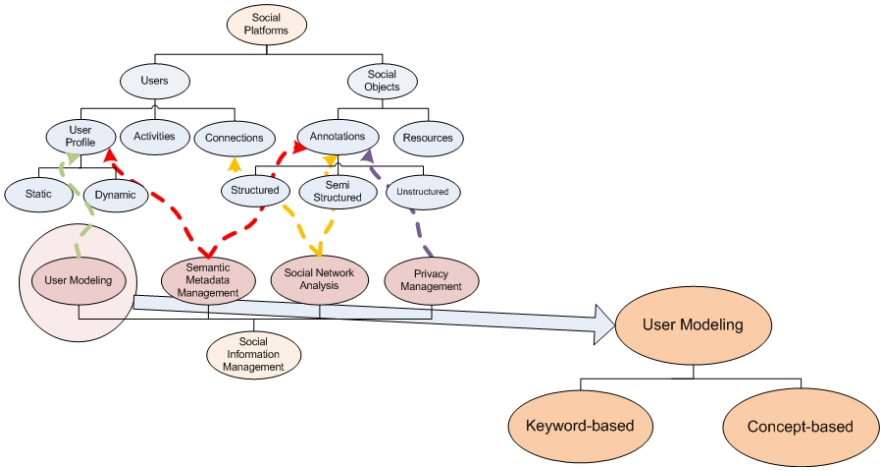


Figure 3.7: User Modeling Strategies: keyword and concept-based

Keyword-based user profiles represent user interests in the form of keywords and do not actually consider their meaning. On the contrary, the concept-based user model associates each keyword in the user profile to an internal or external knowledge base (e.g. ontology, taxonomy or vocabulary). The main role of the knowledge base in this case is to represent relations between concepts and allow their retrieval. However, a difficulty in the case of concept-based user profiles is the association of the keyword to the right concept in the knowledge base (operation called word sense disambiguation in knowledge discovery or semantic linkage in semantic web). This operation is also called semantic disambiguation. Figures 3.7 3.8 gives a high-level overview of the difference between keyword and concept-based user models.

The main advantage of concept-based profiles, as pointed out by [Vallet 2007], is the fact that the resulting profiles will be semantically more expressive. This is mainly due to the fact that a knowledge base contains not only the concepts, but also the relations between them, allowing to perform different expansions. In the case of Social

[25] The other categorization (static vs. dynamic) is considered in Section 3.2

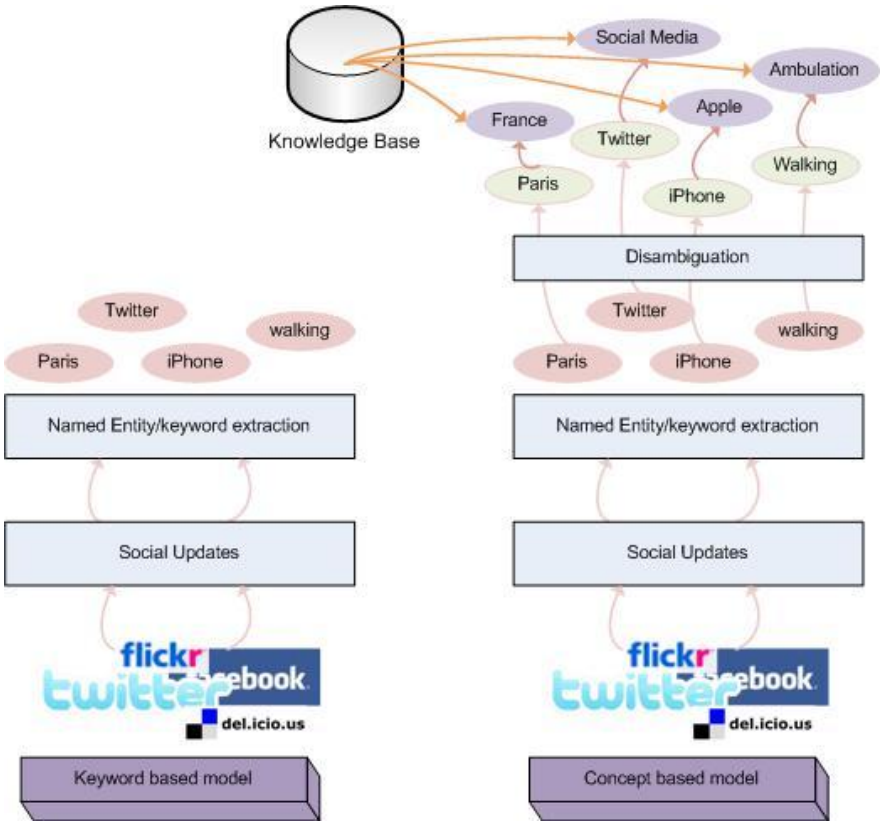


Figure 3.8: User Modeling Strategies: keyword and concept-based

Platforms, this is particularly useful, given the size limitations of annotations (e.g. social tags and status updates). Indeed, social updates are generally short messages and therefore users generally express a summary of their current activity or interest. For example, in a traditional document or blog post, when expressing an opinion about a movie, the user would probably give detailed explanations about the reasons behind the expressed opinion (e.g. bad actors, bad scenery, good special effects etc.). In order to overcome the size limitation in social updates, the semantic expansion thanks to an external knowledge base can be considered a particularly useful operation. Another way of further expanding user interests can be considered the topology of the user's social network. More concretely, in the case of user generated data, content shared by the users's friends can be inserted into the profile. The underlying assumption is that the user shares similar interests and tastes with their friends, in the scope of a better integration of the given community (phenomenon also called homophily) [McPherson 2001a].

The main step in the construction of a user profile is the observation of the user's behavior in a system. Such an observation can consider: (i) the content consumptions of the user and (ii) the content productions of the user. Each observation strategy generally fits a particular ecosystem, which defines the relation between the user and the platform. In both cases, an important step is the computation of weights for the profile concepts, which is generally achieved with adaptations of the tf-idf [McGill 1983] measure used also in traditional information retrieval.

Lots of work has been done in the derivation of user profiles from e.g. web navigation. In this case, the content of the consumed web sites is extracted and injected into the profile. Profile items are weighted using tf-idf or other statistical measures, such as the time spent on the particular web sites containing the given item. In [Sugiyama 2004], the web navigation of the user is monitored for profile construction. The profile is then used to adapt web search results, based e.g. on the derived expertise of the user in topic of the query. [Shahabi 1997] defines a framework for the clustering of similar users, based on their web navigation experience. Users in the formed clusters are then recommended to build a community and engage in discussions based on the viewed web pages.

The second category of user profiling, based on content productions, is more recent. Expertise derivation based on tags [Budura 2009] is a good example of this approach in the enterprise, based on online social applications deployed and interactions performed in them. However, it should be noted that still few work considers the construction of user profiles from social awareness streams [Wagner 2010].

An important issue in the domain of user modeling specific to social platforms is the fact that in these systems users are not willing to spend much time describing their detailed preferences and even more, to share long updates, rich in information. Therefore, in real scenarios, user profiles tend to be very scattered. A possible solution to this issue has been identified by [Vallet 2007] with a semantic preference spreading mechanism which expands the initial set of user preferences stored in user profiles through explicit relations with other concepts in a domain ontology. The approach is based on the so called *Constrained Spreading Activation (CSA)* strategy, introduced and discussed in [Cohen 1987] and [Crestani 2000].

Another recent approach [Michelson 2010] focuses on discovering the topics of interests of a particular Twitter user. The profiles are composed of category terms and therefore are very generalized. The approach to discovering a Twitter user's topic profile hinges upon finding the entities about which a user Tweets, and then determining a common set of high-level categories that covers these entities. An external knowledge base (Wikipedia) is used for entity extraction and disambiguation. All capitalized, non-stopwords are used as possible named entities. This ensures high recall (e.g., we retrieve many possible entities). Once the entities in a Tweet discovered, the system disambiguates them by leveraging Wikipedia as a knowledge-base, by comparing the local

context of the named entity in the Tweet with the text of all candidate wikipedia pages. This approach shows that the discovery of user profiles from tweets is a growing field and can be an input for promising applications. Also, this approach shows the usefulness of including category concepts in a user profile in order to expand the recommendation accuracy (e.g. a user who asks about "Theo Walcott" should be connected to a user who shared about "Arsenal Football Club"). Another interesting conclusion of this work is the necessity of using external knowledge bases in text analysis on microposts for the following main reasons:

- Traditional disambiguation approaches are generally statistical (e.g., using co-occurrences) and require large training corpora. Such a corpora is difficult to be built from microposts.
- Tweets are short, and there are few users who generate enough of them to create a large enough sized corpus for deep analysis.

3.5 Privacy Management in Social Platforms

After the consideration of fields whose contributions are related to the management of one or more pillars, we introduce privacy management, as a cross-field of increasing importance in social platforms (Fig. 3.9).

Privacy of the users and of the content they share is a fundamental problem in a social platform. As mentioned before (Chapter 2, Section 2.1.2), users in a social platform have a profile, activities and connections. These connections belong to multiple social spheres, e.g. friend, family and coworker connections. For this reason, in most social platform users can define the visibility of a shared content. In this section, we examine the current mechanisms proposed in the literature for the management of such visibilities.

A taxonomy for privacy management is proposed by Barker et al., [Barker 2009] in the form of a tuple $P =< p, v, g, r >$, where p stands to the purpose of a data collection (e.g. medical information about a patient), v to the visibility, g to the granularity and r to the retention.

These dimensions have the following meaning:

- purpose: the reason for which a data item is collected.
- visibility: who is permitted to use/access data provided by a provider
- granularity: what level of precision should by shared to a given third-party (e.g. a teacher does not need the same level of precision about the health of a student, as a doctor)

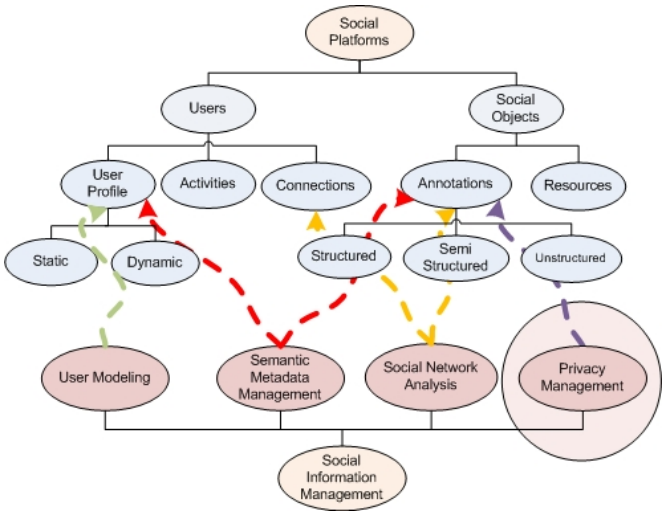


Figure 3.9: Focus on Privacy Management in social platform

- retention: the retention period for the collected data, i.e. how long it can be used.

More clearly, the purpose of the collection is the reason for which the data has been collected for. This dimension contains five degrees: Single, Reuse same, Reuse selected, Reuse any, and any. The visibility defines the actors for whom information should be shown to. Its values are: None, Owner, House, Third party, All World. The granularity as for it defines different variants of the same information that can be used in different situations. Finally, the retention dimension defines the period of time for a piece of information with a specific purpose, a specific visibility, and a given granularity should be conserved in its form. This dimension includes two values: an exact date or unlimited retention.

In a recent work, Carminati et al. [Carminati 2009] proposes an extensible, fine grained access control model to social network based on semantic technologies. Relationships between users in the social network are modeled with various semantic technologies, like OWL and SWRL. Based on these trust relationships, the proposed framework allows to specify rich and extensible policies for authorization, administration and filtering of information.

Besmer et al. [Besmer 2009] performed a user study to highlight potential concerns users have with shared photos in social networks. The main concern of the users refers to close friends or family members who may find them in a photo without their prior agreement. This often occurs on Facebook, where people can tag others in photos, without their permission. An interesting result of the study was the fact that users were not at all concerned that strangers may find their photos or other information.

Another aspect of the privacy problem is that maintaining independent social spheres is a fundamental property of human psychology. In the case of sharing information, we have different preferences in the case of family members, with whom relations can vary from distant to very private and other spheres, like friends, professional colleagues or strangers. Generally, the most important distinction is made between family and other spheres, as we generally try to show a different ego in these cases. [Binder 2009b] performs a study which focuses on the effects of situations in which the content of communication intended to be consumed within a social sphere becomes available in another social sphere. The main result of this study is the fact that the increased visibility of communications in SNS may lead to increased level of social tension. More concretely, in the case of typical social network site use, a user might, for example, post an entertaining message to a friend, only to find that this has negative effects on family relations because of a family member reading a post. This is called the problem of *conflicting social spheres* [Binder 2009c].

The social tension in the case of visibility of communications is due mostly to the fact that the different social spheres are governed by different norms, ethical principles and have different cultural backgrounds. More clearly, people will talk differently to their best friend than to their work colleagues or their family, and also about different things. These important findings suggest that the problem of social sphere management in social networks needs more attention. Therefore, we consider the social spheres of the user as the most important parameter for an efficient privacy preserving mechanism adapted to social platform. This consideration is highly supported by the fact that the current infrastructure of SNS allows the members of one's social spheres (e.g. close family, close friends, school friends, university peers, work colleagues, bosses, work and hobby friend etc.) *to observe communications within spheres other than to which they belong.*

Finally, people tagging for privacy has been introduced in [Razavi 2009] and the idea is to offer a user with a possibility to create dynamic groups of people (according to their tags) which may have different access right levels with respect to their associated tags. The objective here is to make people tagging more dynamic, more trustable, and spam free.

A first conclusion about privacy management is that with the explosion of social networking sites (SNS) and the web 2.0, the traditional privacy management models are not any more appealing. These models generally propose to have access to a given resource according to a social network category. The access follows the traditional allow or deny (i.e. binary) mechanism. We consider that in these environments, the question is not any more of who can access a given information or resource, but what would be the detail level of that information exposed to her. The main reason for this is that the different social spheres that one may have (e.g., Friends, Family, Coworkers, Neighbors, School, etc.) apply also on these digital environments and thus one would naturally like

to expose different levels of detail about resources/information to these social categories and the people composing them. For this reason, it is clear that a more advanced and suitable mechanism is necessary for privacy management in social platform.

3.6 Related Work in Social Search

After the introduction of the main scientific fields related to social platforms and relevant approaches for social information processing in each field, this section will review complete frameworks that tackle the issue of social search.

A traditional search engine performs several steps before it can provide useful and relevant results to the user: (i) the indexation of documents, (ii) the classification of the indexed documents and (iii) ranking of the documents according to a user query. In the case of a social search engine, this changes, as the objective is no longer the retrieval of documents or other kind of content, but social resources, such as people who are somehow related to the user query and thus, can provide recommendations or help the user in the information need expressed in the query.

There are a number of existing frameworks that provide a complete toolkit for the processing of social content together with a corresponding social search strategy. In the following, we will list these frameworks and identify which dimension of social content analysis requires further improvement and also, how could we improve such a fremwork in order to offer users more possibilities to manipulate their profile.

The Flink System An early work in social network analysis that integrates Semantic Web Technologies is the Flink system [Mika 2005], aiming to extract information from different social networks, aggregate it and visualize relations between people. In this system, four different knowledge sources are used: (i) HTML pages for the web, FOAF profiles from semantic web, public data collections and bibliographic data. In each case, specific data extraction and analysis techniques are employed, like co-occurence measures and basic disambiguation of identities using e-mail addresses in FOAF profiles). The aggregated collection of RDF data is stored in an RDF store [26], called Sesame [Broekstra 2002]. This work can be considered as a pioneer of how Semantic Web Technologies can be integrated in the management of social data, even if some important considerations, like privacy management of the extracted user data are missing from the framework. Also, social content is not aggregated into user profiles, which is a limitation compared to the dimensions of users that we presented in Chapter 2.2, Section 2.1.2. However, this work is interesting for our motivation, as the framework helps in understanding the main layers a social systems should implement (e.g. crawling, indexation, presentation etc.). This decomposition of a social search engine into three layers

[26] RDF stores are built on top of standard database management systems, like MySQL or Oracle, with specific structures and tools adapted for the specificities of Semantic Web languages

is of high interest and we consider it as a design pattern for any social search engine implementation.

SocialScope SocialScope [Amer-Yahia 2009b] is a framework that targets the issue of information discovery and management on social content sites. A social content site in this case is similar to our definition of Social Platform. The main novelty of this approach is a flexible and universal algebra, specially adapted for operations on social content graphs. The defined algebraic operations, mostly originating from set theory allow to aggregate social content graphs, to compute unions, intersections and generate different aggregations, depending on the query formulated by the user. SocialScope is a promising approach for the content analysis and data management community. Therefore this algebra is particularly useful for social search engines that target social content retrieval, but not people recommendation. More concretely, this work defines the background for a query language specially adapted to social data.

Social Search Systems Frameworks that specifically target recommendation services based on user profiles are mostly in the category of People Recommendation and Question Answering Systems. Such systems explore either the topology of the network or the content of the exchanges between communities and peers. The main difference to content-based social search is the fact that the result of a recommendation is not a document, but another user or group of users. In this way, the person can interact directly with the recommended user, which provides a more secure and trusted environment for the communication process. Also, such people-to-people interactions are more interesting for the service provider, as they can contribute to the growth of the social platform, which is generally measured by the number of users and connections between them.

[Guy 2009] presents a people recommendation strategy specially adapted for the enterprise ecosystem. The recommendation engine uses information from an organization's Intranet for computing similarity scores between employees. Such information include: (i) paper or patent co-authorship, (ii) commenting of each others blogs or profiles, (iii) mutual connection in other social networks, internal to the organization. Based on an aggregated score computed for each relationship, people are recommended to be added in employee's internal messenger systems. For each recommendation, an explanation is generated, considered an important component of such systems [Herlocker 2000]. A limitation of this approach can be considered the fact that the recommendation only uses statistical information to infer the social proximity between users. More concretely, the content of interactions and exchanges is not taken into account to measure the similarity of interests or information needs. We also mention here the fact that most people recommendation strategies in popular social networks, such as Facebook or Orkut, are also based on this statistical similarity schema.

[Lin 2009] also targets the issue of expertise location in the enterprise environment. The proposed system, SmallBlue [Lin 2009], similarly to [Guy 2009] employs data mining and statistical data analysis techniques to extract profile information for employees. More specifically, the system uses company email as a source of information. Keywords are extracted from each email and a bag-of-words based profile is constructed for employees. An innovative feature of the system is the social explanation of people recommendations, by displaying the social path that connects the user to the recommended person on a specific topic.

[Hannon 2010] goes beyond the previous approach and builds a recommendation strategy using the content of interactions (e.g. status updates) as input. Designed for recommending people to follow in Twitter, the *Twittomender* system allows users to expand their network by connecting to people that they don't know directly, but with whom they share similar interests. Each user in the system is represented by a vector, comprised of terms extracted from their shared messages. A kind of *social expansion* of this basic profile is performed, by taking into account messages shared by people connected to the user. This is based on the observation that connected people share close interest. The computation of profile similarities is achieved by the traditional tf-idf weighting schema in information retrieval and cosine similarity. The Twittomender system is original and different from existing collaborative filtering approaches, as it takes into account to structure of the underlying social network to better approximate the interests of the user. It is however a considerable limitation in the system that no disambiguation or semantic expansion of profile terms are considered. More concretely, the user profile is composed of keywords that might have multiple meanings and this could be a considerable drawback for the relevance of recommendations.

A new generation of social search engines is represented by so-called *Question Answering Systems*. The main difference to the previous approaches is the fact that in this case the system builds a user profile from some kind of user activity (content production or consumption) and uses it to match them with a question formulated by another user.

Aardvark [Horowitz 2010] is certainly the most promising social search engine. Aardvark introduced several innovations in the field of social search. First of all, it is the first system that models the users based on their generated content. For this reason, users provide topics of interest to the system when they subscribe. Then, a crawler extracts further topics from the user's profiles and status updates in social platforms to expand the initially entered profile items. The extraction of topics from social updates is achieved by linear classifiers, such as Support Vector Machine and probabilistic classifiers. Aardvark is not built on top of existing social platforms and lacks a global approach for conceptualizing user profiles.

Another recent social search engine (i.e. CQA) is proposed by [Li 2010]. In this case, the objective is similar to that of Aardvark: route a question to the right person in a community of answerers. This paper introduces two important dimensions for

such systems: (i) the consideration of the answerer's availability and (ii) the question of quality of answers. The quality of answers is estimated by taking into account statistical information about the length of the answer, the time the user took to send it and the feedback of other users. In the case of availability, the system monitors the user's logins and performs a prediction of whether the user will be available at a specific time and date in the future.

We consider a set of criteria to compare the previously described social search systems. The defined criteria, similarly to the comparison of semantic metadata management models will consider how well the social search engine takes into account the dimensions of the user and user generated content in social platform:

- *Social Content Aggregation (SCA).* This criteria rates the capacity of the framework to aggregate social data from several sources.
- *Static Profile (SP).* This criteria analyses if the profile constructed for users is static or updated according to the users' interactions. If the profile is static, this criteria is marked with a +, otherwise -.
- *Dynamic Profile (DP).* This criteria analyses if the profile constructed for users is static or updated according to the users' interactions. If the profile is static, this criteria is marked with a -, otherwise +.
- *Keyword-based Profile (KP).* This criteria considers weather the items composing the user profile are keywords.
- *Concept-based Profile (CP).* This criteria considers weather the items composing the user profile are semantic concepts (i.e. keywords linked to a semantic knowledge base).
- *Semantic Metamodels (SWT).* This criteria defines the degree to which the given system integrates Semantic Web Technologies (e.g. representation or analysis level).
- *Social Expansion of the Profile (SOP).* This criteria considers the social expansion of the profiles constructed in the system. A social expansion takes into account the social network of the user to inject items into their profile, by taking into account the fact that friends share similar interests.
- *Semantic Expansion of the profile (SEP).* Opposed to social expansion, semantic expansion uses an external knowledge base to perform an expansion of the concept that is injected into the profile. As an example, a user interested in a movie, say Gran Torino, could also be interested by its actors, places and other related concepts.

Table 3.2: Summary of the considered dimensions and the affected scores for each system regarding the analysis dimensions

	SCA	*SP*	*DP*	*KP*	*CP*	*SWT*	*SOP*	*SEP*	*RE*
Flink System	+	-	-	-	-	×	+	-	+
SocialScope System	×	+	-	+	-	-	-	-	-
Twittommender System	-	+	+	+	-	-	×	-	-
IBM Enterprise System	-	+	+	-	+	-	-	-	×
CQA System	-	+	-	-	-	-	-	-	
Tag-based Expertise profiles	-	+	-	-	-	-	-	-	-
SmallBlue System	-	-	+	-	+	-	-	-	×
Aardvark System	+	+	×	+	+	-	-	-	-

- *Recommendation Explanation (RE).* This criteria takes into account if the system generates some kind of explanation to the result.

Table 3.2 provides a summary of the analysis where columns represent the different considered models and rows the considered analysis criteria (i.e. dimensions). The following notations have been used in the table: (i) columns represent the semantic user and content management models, the comparison criteria, (ii) each cell of the table contains a rating: "+" sign means that the given model provides/supports a considered criteria, "×" means that the considered criteria is the (one of the) most prevalent characteristic of the corresponding model and can eventually be considered as a reference for the given criteria. On the contrary, the "-" means that the given model does not deal with the considered criteria at all. This should not be considered automatically and uniquely as a lack of a particular model but rather as simply the non consideration aspect.

This comparative table shows that in current social search systems, the issue of recommendation explanation is still not well tackled (which is also strongly related to privacy managment). Also, few frameworks benefit from Semantic Web technologies on a data storage or data enrichment level.

3.7 Discussion on Social Information Processing and Social Search

In this chapter we revisited approaches to social information processing techniques that are specially adapted to the building blocks of a social platform: (i) *Social Network Analysis*, (ii) *Semantic Metadata Management* and (iii) *User Modeling*. After this,

reviewed several approaches in *privacy management*, a fundamental activity for the protection and sharing of personal private information. Finally, we presented the most important social search frameworks in order to understand their objective and design principles.

We first introduced the field of *Social Network Analysis* in order to have an upper-view of the main techniques used to understand the processing of data shared in these systems both from a structural and semantic perspective. To our challenge, as defined in the introduction, clearly, the semantic data processing is more closely related. More specifically we identified works that have the objective to transform unstructured social data into hierarchical vocabularies ([Brooks 2006] [Zhou 2007]) and identified algorithms ([**?**]) that allow to rank social content. They are adaptations of the PageRank algorithm and take into account additional social parameters in order to retrieve tags that are most influential in the platform, by considering e.g. the number of connections the owner of the tag has. However, these approaches do not take into account the meaning of a specific tag. Therefore, more relevant to our work are algorithms developed by [Angeletou 2008], which associate to tags a corresponding concept from an ontology in order to approximte the meaning and enrich it with neighboring concepts. This work proves that it is possible to automatically enrich folksonomy tagsets with ontological entities. We reuse this result for our toolkit of the processing of microposts. It is also clear that the process for the profile construction should also be as automatic as possible, but there are several parts which will require user intervention in the form of relevance feedback.

As for as *Semantic Metadata Management models*, their substantial review helped to better understand the main dimensions of social data both in case of user profiles and in the case of social web models. The most interesting model to be reused for storing social data in our case is MOAT, as it integrates the meaning dimension for tags and can be easily extended with other additional contextual information with regards to tags. Since the conceptual model of MOAT can be reused without modifications for our purpose, it is out of the scope of this work to detail how the semantic datastore is implemented in the framework.

The most relevant works in the field of *User Modeling* are that of [Szomszor 2008] and [Cantador 2008]. The most interesting finding in [Szomszor 2008] is the fact that the use of an ontology will allow recommendation systems to find out how specific the user interest is, and use this information to fine tune recommendations. Inferring interests by analysing links or paths in the ontology can help uncover implicit interests. We will reuse this result in our profile building mechanism. [Cantador 2008] continues this work with a formal model. In this case the matching of tags to category concepts is achieved by a morphologic matching between the name and the categories of the entry, and the names of the ontology classes. The ontology classes with most similar names to the name and categories of the entry are chosen. This approach is also relevant to our

objective, however their matching process does not benefit from the textual description of ontological concepts to improve precision.

Another important issue is the fact that, once we have the explicit user profiles, how to propagate the interests scores to expanded concepts? The spreading activation algorithm introduced by [Crestani 2000] is a significant contribution to this purpose, but does not take into account the names of properties that connect two concepts in the knowledge base. More specifically and in a none-limitative example, a propagation to generalization concepts may be more important as to other neighboring concepts.

As for as privacy management, it can be observed that although there are some early attempts to provide social platforms with a more granular privacy managment strategy (e.g. [Barker 2009] [Carminati 2009]). There is a need to further elaborate on this topic and integrate a generic approach to social platforms as this can significantly improve the way social data is shared, exchanged and reused. More specifically, a granular approach could be used to generate explanations to recommendations, by computing what level of granularity to show for a specific user profile attribute.

The construction of a social search framework is not a new challenge, as the list of reviewed systems shows (e.g. Aardvark, Twittomender etc.). The review of social search systems has been proposed to clarify for each of them their main characteristics. The conclusion of this review is that our proposal has already been partly adressed by frameworks such as Aardvark (people recommendation for a question), Twittomender (the use of Twitter as an input platform) and Fink (a three-tier architecture). However, we observe that these systems are not specifically tailored to build user profiles from microposts (i.e. take into account the fact that content is short in size and poor in knowledge) and consider the style of these posts to understand user expertise or infer interests. Also, each framework has limits with regards to our objective: Fink is not centered on user profiles, Twittomender does not leverage semantic knowledge bases in the profile construction and Aardvark does not allow to leverage an existing social network of the user.

In conclusion, based on our analysis, we define two main challenges for efficient social information management in social platforms:

- Conceptual User Modeling in a Social Platform. What is the right strategy to model the user in such a system? How can we build such a user model based on shared content? How to provide more expressivity to such user models by leveraging Semantic Web knowledge bases? How to score concepts in the user profiels in order to identify users who may be intersting for a given topic?
- Privacy Management in social platforms. What could be the right strategy for the privacy management of users in a Social Platform?

The framework we present in the next section provides a complete architecture for transforming shared content into conceptual user profiles. We introduce an algorithmic

toolkit for this process and show the different steps required to structure social data in user profiles.

Part II

Contribution, Evaluation and Conclusion

Chapter 4

A Semantic Framework For Social Search

Contents

After the upper-view of related fields (i.e. Social Network Analysis, Semantic Metadata Management, User Modeling, Privacy Management), the review of relevant works to information management, discovery and processing in social content sites and previous work in the more specific area of Social Search, this chapter will present our solution for a framework that implements a more advanced social information management strategy in social platforms.

4.1 Introduction

In a traditional web search engine, the challenge lies in finding the document that best satisfies the information need of a user. A social search engine goes beyond this approach, by suggesting users who may know the answer to the information need or directly showing the answer of a human, step which involves in a first time the identification of this user.

In order to present our approach, we first analyse more thoroughly microposts and communities in such platforms with the objective to have a clear understanding of their structure and style. Thereafter, we introduce our framework and present each component that contributes to the social search strategy.

More specifically, this chapter is organized as follows:

- in the first section we present a short study of an analysis of a Twitter community and microposts (structure and semantics) in order to understand their general style and organization and identify challenges for their semantic processing (i.e. how to extract meaningful information from a micropost?);
- the second section will present an upper-view of the toolkit composed of algorithms for the user expertise profile construction;
- in the following sections, we inform on the framework and a corresponding proof-of-concept application that we designed and implemented to perform social search on top of Twitter, one of the most popular and dynamic social platforms. The reason why we selected Twitter is multifold: (i) people in Twitter have a large diversity of connections, from friends, family to coworkers and even strangers; (ii) connections in Twitter are not mutual; and (iii) in Twitter, shared messages have a large diversity in terms of content.

4.2 Analysis of Microposts: the case of Twitter

As mentioned before in this work, posts shared in social platforms are different from traditional documents, mainly because of an additional social dimension that motivates people to share more frequently short snapshots of their lives. It is therefore important for our framework to have a basic understanding of the structure and semantics of such posts and also of the structural properties that govern communities in such systems. We first studied several aspects of a user community in Twitter in order to gain more insight into its topology.

One of the most obviously visible indicators of a user in a social platform is the number of friends, corresponding to the number of other people who can see the contributions that they publish.

In many social platforms (including Facebook), this relationship is symmetric-both parties must accept the relationship (usually one proposing the friendship and the other accepting). This is equivalent to a non-directed edge between two vertices in a graph.

In other social platforms (such as Twitter), the relationship is asymmetric. In Twitter terminology, a user can add any other user to their friends list in order to subscribe to that users updates. A user's followers list contains the other users that have subscribed to their updates. These are equivalent to outgoing and incoming directed edges in a social network graph.

4.2.1 Origina of Twitter Sample Data

The sample data utilized for the analysis is a subset of our collection of posts from Twitter. The 290735 rows (each representing a Twitter user) were saved to a comma-delimited file and then imported into an R data frame [1] for analysis. First of all, for each user we collected the number of followers and friends.

4.2.2 Followers Analysis

One of the most important indicators of a Twitter user's importance in the social network is simply how many other users have chosen to follow them.

.

For readability, the R statement used to generate the histogram in Figure 4.1 has the range limited to those users with 2500 or fewer followers which includes 86% of the users sampled. To put the figure in perspective, if 100% of the users were represented at the same scale, Figure 1 would need to be almost 3000 times wider. There were 5161 sampled users (1.8%) who don't have any followers, and 31903 sampled users (11.0%) have 10 or fewer followers.

[1] http://cran.r-project.org/

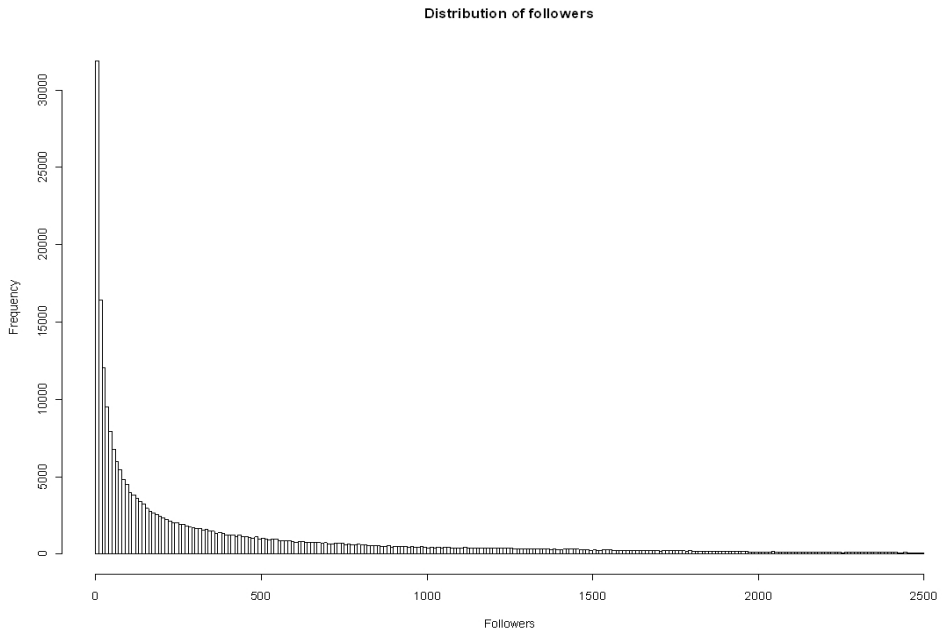


Figure 4.1: Distribution of followers in Twitter [2]

Figure 4.1 shows the ***long tail (power-law) distribution of followers among the sampled users.***

4.2.3 Following Analysis

A Twitter user $u1$ can "follow" another user $u2$. The user $u2$ is then considered a friend of $u1$, while $u1$ is then considered a follower of $u2$. The messages of $u2$ appear in the friends feed of $u1$, and $u2$ may be notified of the relationship by email. The relationship is not reciprocal or symmetric-no permission is necessary to declare another user your friend and become their follower. The other user may (or may not) independently declare you their friend and become your follower [3]

For readability, the R statement used to generate the histogram in Figure 4.2 has the range limited to those users with 2500 or fewer followers which includes 93% of the users sampled. To put the figure in perspective, if 100% of the users were represented at the same scale, Figure 4.2 would need to be 280 times wider.

As a histogram, Figure 4.2 counts users by buckets of 10 (i.e., the first bar indicates that there are 16989 users following 0-9 other users). The exact counts for the first

[3]Details about the construction of the graph in R of Figure 4.2: hist(twitter$friends$[$twitter$friends < 2500], breaks=seq(-0.5,2500.5,by=10), main="Distribution of following", xlab="Following", freq=TRUE, xlim=c(0,2500)).

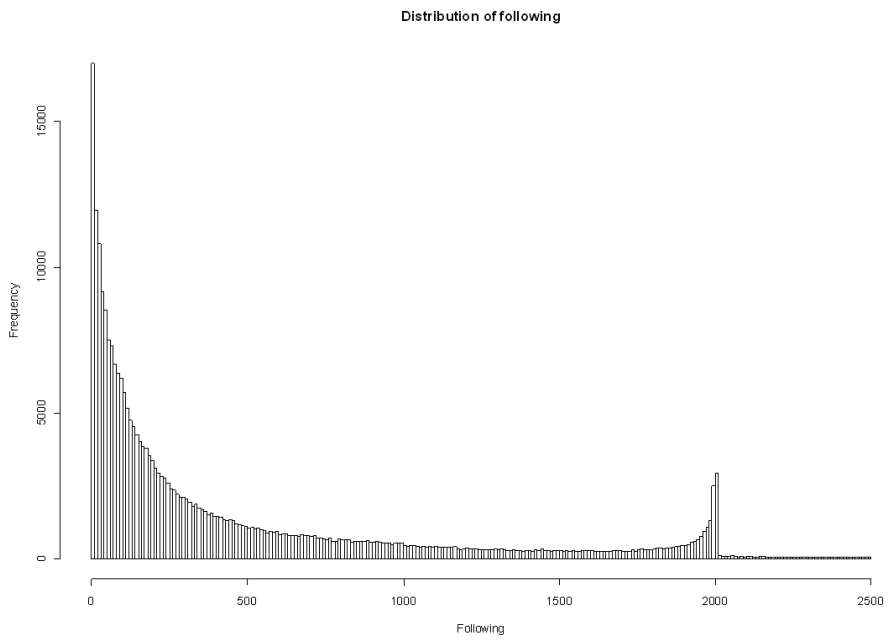


Figure 4.2: Distribution of following

quartile (equivalent to the leftmost 1/35th of Figure 4.3) are shown in Figure 4.3. Almost 1.2% of the sampled users are not following anyone, with 0.8% following exactly 1 other user [4].

The spike in users with around 2000 followers (shown with more detail in Figure 4.4) is a consequence of the limits put in place to prevent "follow SPAM" put in place. In order to a user from indiscriminately following as many other users as possible, Twitter policy limits them to "following 2000" other users, or to "following 110%" the number of users following them (whichever is greater). Every sampled user to the right of the spike must either have more than 1800 followers (and is able to follow 110% times that many), or must have been following more than 2000 before the policy was put in place in August 2008.

Note that, for some reason, the "following 2000" spike actually appears to be at 2001. It appears that there is a trend for some users to follow as many other users as the system will let them, and would probably follow more if permitted.

[4] Details about the construction of the graph in R of Figure 4.3: hist(twitter$friends$[$twitter$friends$ < 70], breaks=seq(-0.5,70.5,by=1), main="Distribution of following (first quartile)", xlab="Following", freq=TRUE, xlim=c(0,70)).

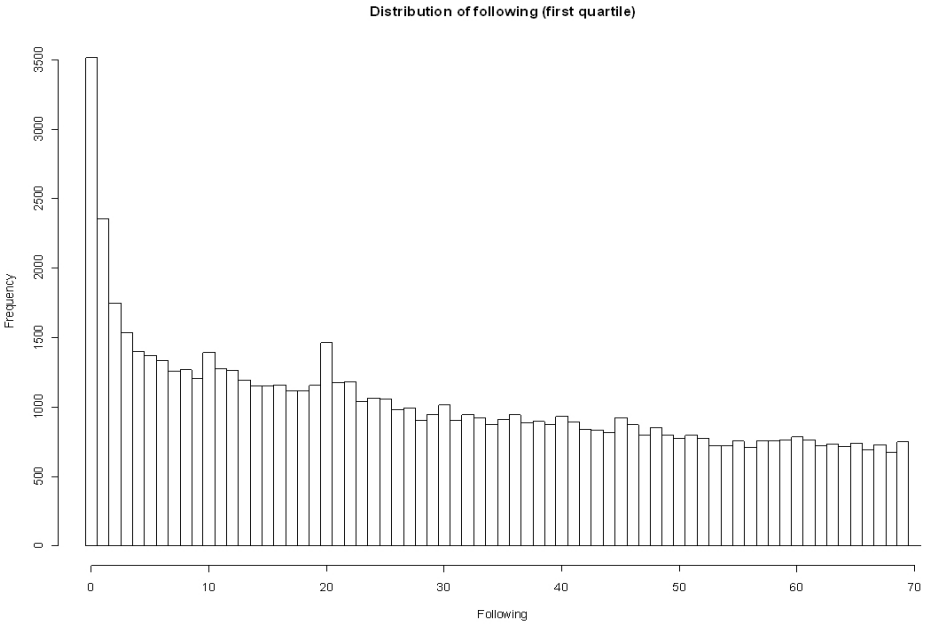


Figure 4.3: Distribution of following (first quartile)

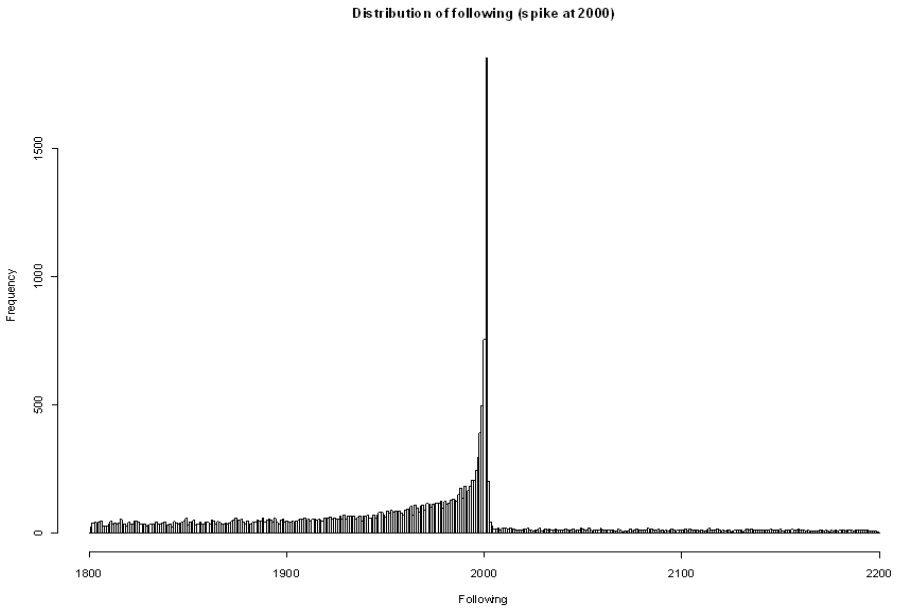


Figure 4.4: Distribution of following (spike at 2000) [5]

4.2.4 Structure and Semantics of Microposts

Our objective in the following is to compare the kind of information shared in the messages and understand their distribution. A message containing an URL is generally an interesting finding that the user intends to share with their community. Generally, such a message is composed of an URL and some additional tags that summarize it. Another kind of message is that of personal activities, where the user just shares what she is currently doing, a question or other kind of information about their current context. Such messages contain generally words that are characteristic for the user's small community (e.g. friends, family) and that contain personal information and expressions.

First, our intention was to extract tags from microposts and understand their distribution. For two independent social update sets, composed of both 40 000 messages, we both obtained power-law distributions of the entities in the posts. Note that for this part of the analysis, we employed Matlab instead of the R framework. Therefore, the graphs are slightly different in design.

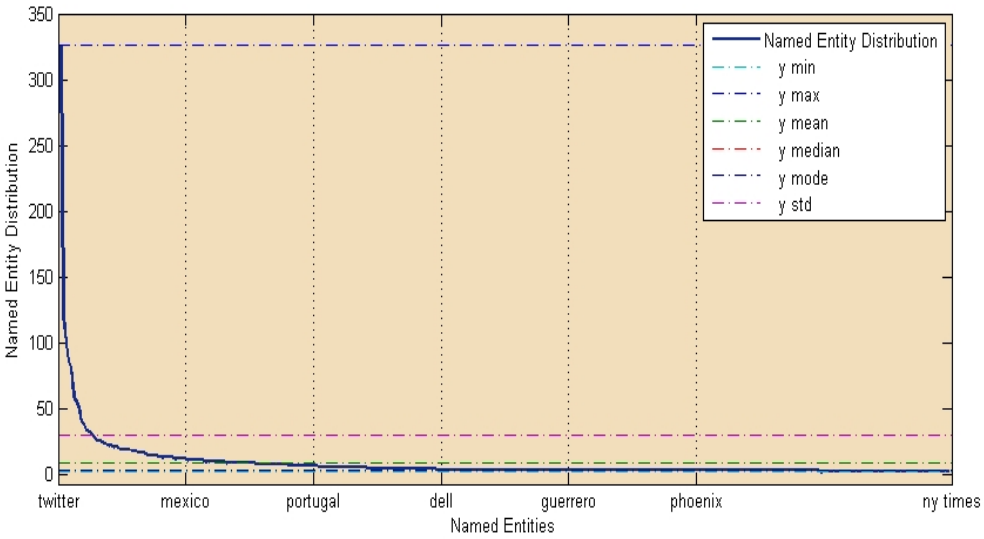


Figure 4.5: Distribution of tags in social updates for a subset of users who follow the "CNN" account

Finally, our objective was to understand the semantic structure of social updates. More concretely, we performed entity and keyword extraction on the message set in order to observe weather social updates are composed of named entities, keywords or both and in what proportion. In this case we considered an URL also as an entity, as the corresponding web page is composed of a set of entities and keywords.

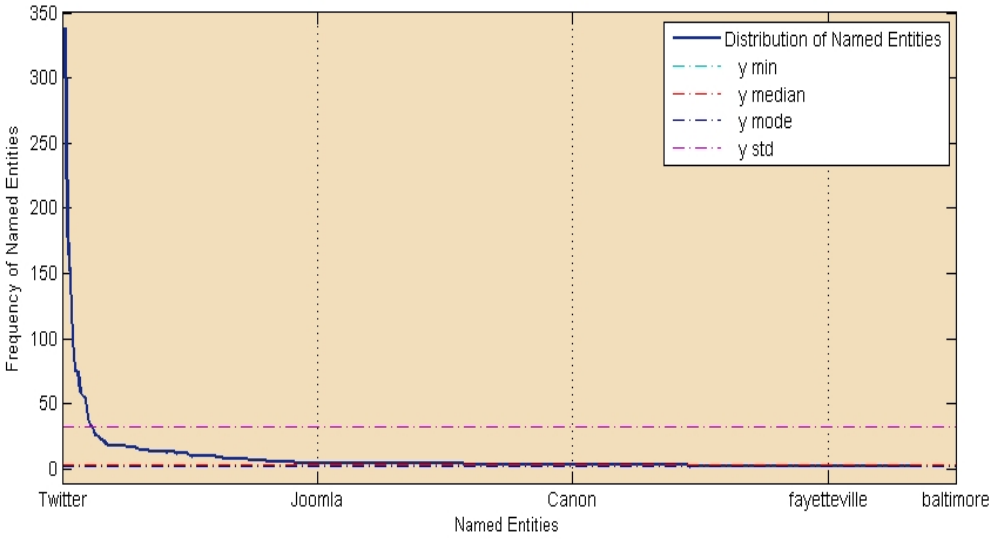


Figure 4.6: Distribution of tags in social updates for a subset of users who follow the account of user "jstan"

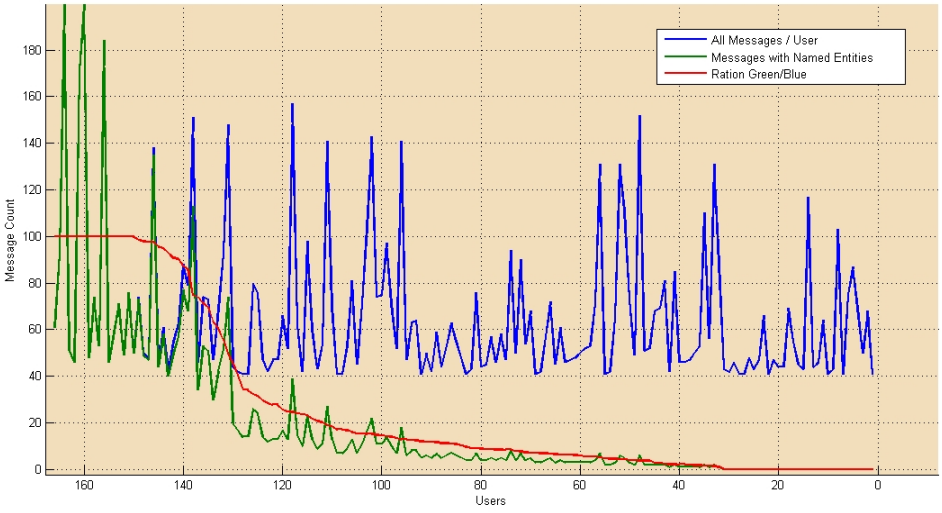


Figure 4.7: Semantic Structure of Social Updates in a Twitter community

4.2.5 Discussion on the Analysis of Microposts

This analysis shows that in Twitter there exist different user segments. Some users share lots of web resources, others more about personal activities, which is a similar finding to that of [Naaman 2010a]. Users who share such messages are part of smaller communities, like a family or a small community of friends who uses Twitter to communicate their current activity and plans. On the contrary users who share web findings and less private messages are generally part of bigger communities where there is a combination of people from different social spheres. Therefore messages must reflect information that is interesting for all, therefore personal context updates are generally omitted in this second case.

A general conclusion of this analysis of two different user segments in Twitter is the fact that the topic of shared messages follows a power-law distribution, i.e. a law characteristic of the topology of social platform communities. However, this is the first study that proves that the topic of short messages from multiple users also follows this rule. Also, the analysis of these different user segments in Twitter show that people share extremely heterogeneous content that includes hyperlinks, personal status activity updates, etc. Some users use it for sharing context in small family or friend communities, whilst others use it for self presentation and attention attraction in bigger communities.

A more specific conclusion deals with the different user categories. As it can be observed, three categories of users can be distinguished. A first category is composed of users who share messages containing only keywords, probably for communicating their activity in a very general way (such as "I like painting"). The second category includes users who share more precise and specific information dominated by the presence of named entities or URLs (such as "I like *Claude Monet*"). Finally, there is a third category doing both.

The strategy to capture user interests from shared content must therefore be adapted to all categories of users and this is our key argument for integrating Linked Data in the approach. Indeed, different operations in Linked Data graph allow to explore the semantic neighborhood of concepts and therefore to expand a given concept with either more general or more specific ones, depending on the category to which the user best corresponds. In other words, the exploration of semantic neighborhoods of concepts in Linked Data may allow to better approximate the meaning of very general posts, as well as to generalize posts of users who share very specific information.

An interesting point in this study appears also the power-law distribution of both connections and topics of shared content. An immediate consequence of the fact that power-law governs in these platforms with regards to content distribution and connections (note that the same property was demonstrated for social tagging systems by [Heymann 2006]) indicates the fact that it may be interesting to decompose such communities into smaller ones centered on specific topics of discussion, where users can discover more easily people with similar expertise or interests. In fact, because of

the power-law, it may be difficult for users to discover people who might be experts in their information needs, but are newcomers to the platform or have different connection habits. Therefore, in order to identify interesting people, it might be more interesting to consider the semantics of interactions, instead of e.g. people with most connections who share on a given topic.

4.3 Framework of the Social Search Engine

After the study of the topology of user communities in Twitter and content sharing habits of members, in this section we introduce the main principles and architectural view of our framework for social search. The structure of this section is the following:

- in the first part, we present an upper-view of the framework and emphasize on its main layers
- in the second part, we focus on each individual layer and corresponding components, including algorithms in the frame of social information processing necessary for the construction of user expertise profiles
- in the third part, we focus on the important issue of privacy management in social platforms and present a generic granular approach

Figure 4.8 shows the different layers of the framework for:

- The *Content Capture* Layer is dedicated to the capture of social information (e.g. interactions, posts) from one or more social platforms
- The *Analysis Layer* implements specific algorithms for the analysis on microposts. The objective of this layer is to transform microposts into meaningful profiles for each member, that reflect their expertise and interactivity.
- The *Knowledge Layer* provides the interface to third-party applications that intend to implement social search strategies on top of the framework. The description of this layer is out of the scope of this work. The review of semantic metadata management models may help in choosing what models to use in this layer.

Figure 4.9 shows the organization and relationships between said layers.

The structure of the database for the content capture layer is as follows. Specific tables store information about the user and interactions. We consider in the following the interaction part of said database structure. Figure 4.10 depicts the tables and internal relations. An interaction can have one or more participants, where each participant is a user. An interaction has a type, which can be an element from a predefined set

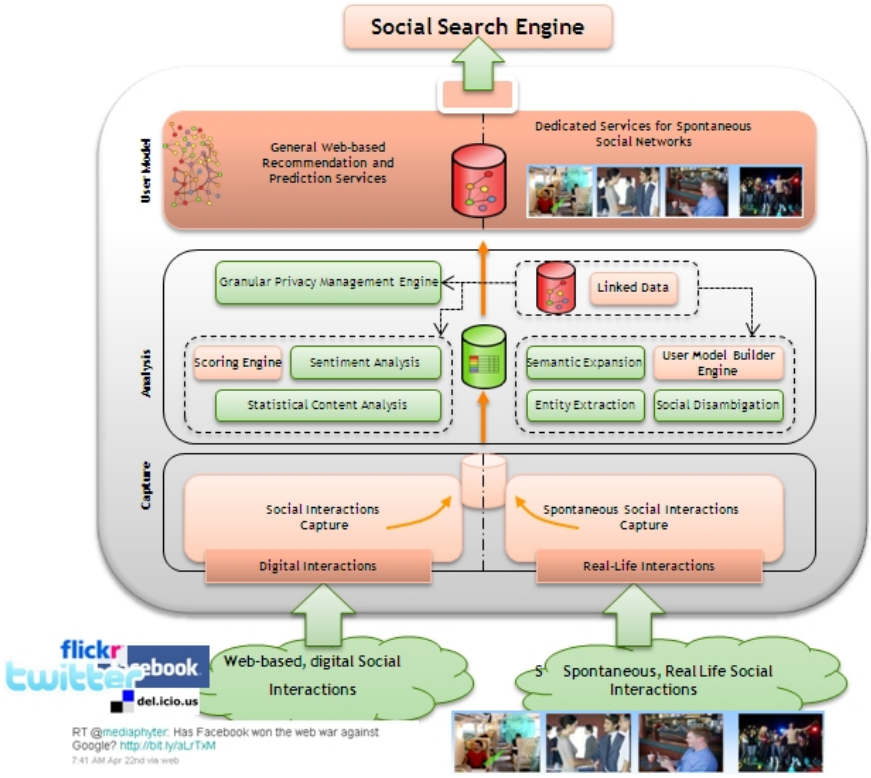


Figure 4.8: Semantic Framework for Social Search

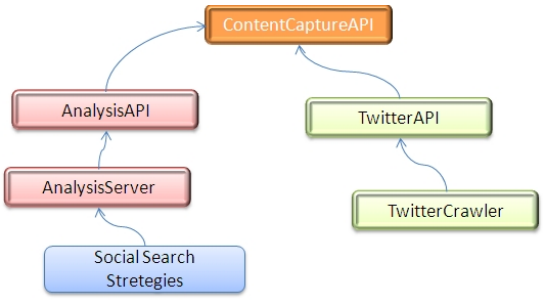


Figure 4.9: Organization of layers in the Semantic Framework for Social Search

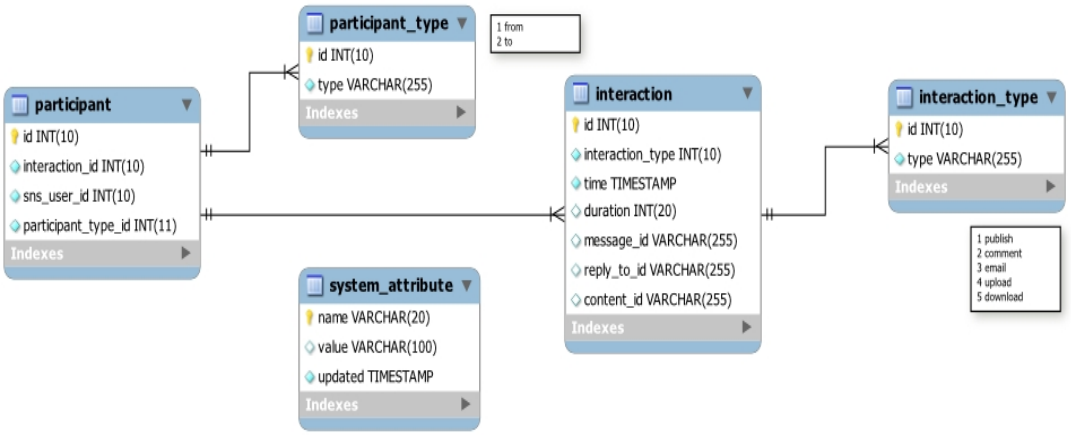


Figure 4.10: Structure of an interaction

that contains the most common interaction types in a social platform, such as the publication of a message, a comment to an already existing message, an upload of a media object etc. Additional metadata about interactions are also included, such as duration, timestamp and length.

A detailed example of the technical implementation of this layer can be found in the description of a proof-of-concept implementation of the framework. For this, detailed information can be found in Chapter 5.

In the sequel, we focus on the most important part of this framework, the Analysis Layer and on the description of the underlying algorithms for the analysis of microposts.

4.3.1 Analysis Layer: Toolkit for User Expertise Profile Construction from Microposts

This section presents the mechanism of user profile construction from the microposts shared in the platform. These profiles give a representation of the user in the framework, based on the content produced in a social platform.

In our case, the main differences between the user profiles we consider and profile models in the literature introduced in Section 3.4 are the following: (i) we construct the profiles from the content productions, not consumptions and (ii) our intention is to model expertise, but also interactivity, i.e. the motivation to conversate about a given topic. In other words, we intend to find a balance between expertise and motivation to interact, in order to retrieve social resources, such as people who may give high-quality answers to questions and thus reduce the knowledge latency in a human organization.

Therefore the source of data used to profile users is different from traditional ap-

proaches of profiling: instead of being interested in consumed content, such as web navigation, we base our profiling approach on social awareness streams as this is the most popular form of content production and in this way, users share their explicit interests. In order to build these profiles, we need to perform several types of analysis on the microposts (Figure 4.11).

More concretely, the profile construction is composed of two major parts: (i) the identification of semantic data that will compose the profile and (ii) the scoring of said semantic data, which will allow to differentiate between users' level of expertise and interactivity for a given question.

These two major parts of the Analysis Layer are composed of the following atomic software modules, each representing a specific step in the profile construction and update process:

- *the transformation of microposts topics into meaningful concepts, linked to semantic concepts from a well-defined Linked Open Data knowledge base. This is achieved by the following two algorithms:*
  - the extraction of keywords and named entities from the micropost and the semantic matching of such extracted content to corresponding Linked Data concepts. This operation is performed by our Social Semantic Matching algorithm (SoSeM).
  - the semantic expansion of concepts, by leveraging links between said semantic concepts in the associated Linked Data knowledge base. This operation is performed by our Semantic Expansion algorithm (SemEx).
- *scoring of concepts in the user interaction profile.* Algorithms in this category associate progressively to each concept in the user profile a series of scores that measure the degree of expertise and interactivity.

The question therefore is how to measure expertise and interactivity, based on the style the user shares about a given subject area?

In order to measure this, we define a set of scores that combined together reflect not only the expertise of the user, but also the cognitive processes he/she had when performing the content production task. Indeed, it is one of the biggest limitation of current approaches for people recommendation systems, that they don't take into account the fact that the underlying entity of the information retrieval process is a human and not a document anymore. Humans have a series of cognitive processes that might significantly influence their motivation to answer question or to interact with others in a given subject area. These can be the influence the given object had on the user for example, which can be measured by analyzing the sentiment polarity of the messages. Also, we consider other statistical scores in order to identify the degree of

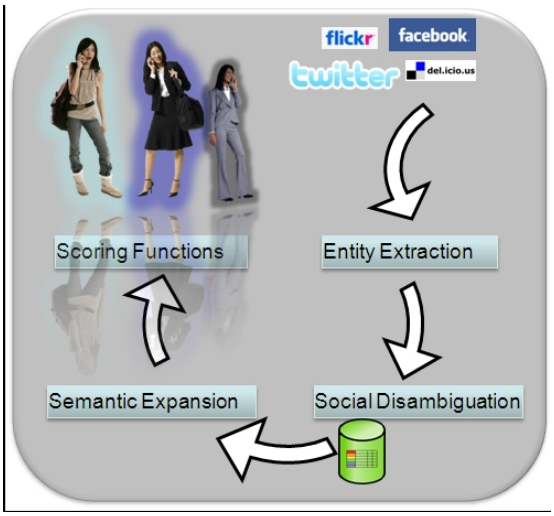


Figure 4.11: General Process to transform microposts into user profiles

expertise, interactivity of the user, such as the complexity of the shared message and the temporal distribution of messages that contain a given keyword. The underlying assumption is the fact that people who share regularly and longer, richer messages in a given field can be trusted more, as it may belong to their job or a hobby.

In conclusion, we will consider the following scoring functions to identify and measure expertise and interactivity:

- **Tf-idf**. This is the standard weighting schema in information retrieval that allows to score concepts that best reflect interests and distinguish a user from others.

- **Sentiment polarity analysis**. The sentiment expressed in a micropost is a primary indication of how profoundly the object of the message influenced the user. Therefore, sentiment can also reflect interactivity. Users who share messages with either high or low sentiment polarity are more interactive and more interesting to recommend, as those who shared neutral messages.

- **Complexity of microposts**. Entropy allows to measure the complexity of a micropost, and correspondingly, the diversity and richness of the vocabulary used by the user to talk about a given topic. We consider that microposts with high entropy reflect the intention of the user to share a high quality information.

In the following, we consider each component of the interaction profile construction process and present the corresponding algorithms.

4.3.2 Generic User Profile Model

The formal model of the user profile is the following. A user profile is composed of atomic user profile items, each having a specific weight.

We consider U the domain of all users involved in the social platform. CU represents the set of items correlated with user u, i.e. $CU = c|P(c,u) > 0$. Therefore, user u and item c are correlated when $P(c,u) > 0$, P being the weight of the item in the profile.

Each profile item is an entity (keyword, named entity) extracted from at least one content production of the user and connected to at least one semantic concept present in at least one semantic knowledge base.

The main arguments for this choice is that this kind of representation is richer and less ambiguous than a keyword-based or item-based model. It provides an adequate grounding for the representation of coarse to fine-grained user interests. A semantic knowledge base provides further formal, computer-processable meaning on the concepts (who is coaching a team, an actor's filmography, financial data on a stock), and makes it available for the system to take advantage of.

Knowledge base-originated semantic concepts are more precise, and reduce the effect of the ambiguity caused by simple keyword terms. For instance, if a user states a message containing term "paris", the system does not have further information to distinguish Paris, the french capital, from Paris, a person or product.

This representation has several advantages to a keyword-based approach:

- Parents, ancestors, children and descendants of a concept give valuable information about the semantics of the concept
- Apart from hierarchical properties, other arbitrary semantic relations can be exploited for the expansion of concepts

In the following we focus on the two most important steps in the user profile construction process: (i) how to transform the keywords into semantic concepts (called Semantic Matching) and (ii) how to perform the expansion of concepts using the knowledge base (called Semantic Expansion) and the propagation of weights to expanded concepts?

4.3.3 Semantic Matching

As we mentioned before, we focus on using Linked Data knowledge bases (more precisely DBpedia [Lehmann 2009]) to transform user's contents into concepts in order to benefit from the rich semantic data that can be extracted from it for the description of a concept. The main difficulty in connecting named entities and keywords extracted from contents shared by the user to Linked Data concepts is the choice of the best concept(s) from the knowledge base that best approximates the intent of the user.

The Figure 4.12 shows an example of the named entity "Apple" that can refer to both concepts "Apple Inc." (standing for a company) and concept "Apple" (standing for a fruit). Thus, we need to tackle the issue of disambiguation in order to associate the right meaning to the semantic concept that will be included in the user profile. The idea originates from the cognitive process during our natural language conversation. Normally in a conversation, we depend essentially on the context of the conversation to disambiguate a word. Similarly, in order to associate keywords or entities in a social update to the right concept in Linked Data, contextual cues are necessary to allow restricting the semantic field of the social update. In traditional documents, generally there are sufficient contextual cues to overcome such ambiguous situations, where the meaning of a term is not straightforward.

In our case, the short nature of posts requires to find these cues elsewhere, so in our algorithm we consider two main additional sources of contextual cues:

- The first contextual cue is user-related, which consists in building incrementally a *vocabulary* from all social updates of the user. The assumption behind this first additional context is that there is a probability that the user previously shared some content in a related semantic field (e.g. a user who posted about "Apple" might have shared before about other Apple products, such as the "iPhone").

- The second additional contextual cue is community-related. On social platforms users are members of different communities, which influence each other in terms of interests. Users participate to a group or a community because he/she is interested in the community' interests and as a consequence of this participation, users have intention of using commonly known keywords to make his/her contents easily understandable by the community. This second contextual cue is used only if the user-related is not yet available or not sufficiently rich (e.g. user has shared few messages, but has lots of friend connection). More specifically, it is a solution for the so-called cold-start situation and consists of aggregating the most recent messages of friends connected to the user and constructing a vocabulary from the content of these messages.

In the sequel, we present the inclusion of the first contextual cue (i.e. user-related) in the semantic matching process. The inclusion of the second contextual cue (e.g. community-related) is based on the same algorithmic principles.

4.3.3.1 Social Semantic Matching Algorithm (SoSeM)

After the introduction of the general principles of the semantic matching process, in this section we present the main steps of the algorithm. To recall, the work needed to be done here is for each relevant keyword found in a message, the algorithm must look

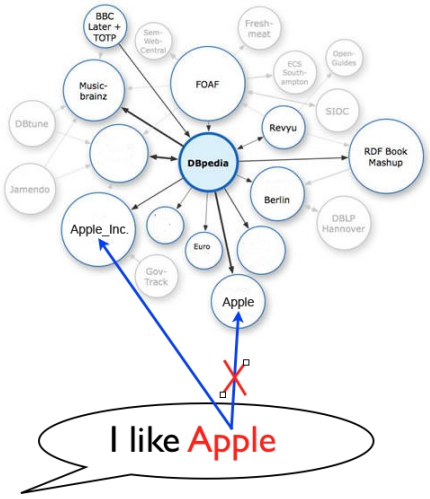


Figure 4.12: Matching from content to Linked Data concept. Example of how a keyword can refer to more than one concepts

into a Linked Data knowledge base (e.g. DBpedia), search for related concepts and return the ones that approximate at best its meaning.

Generally speaking, user interests can be represented by keywords. Hence, their extraction is essential in user interests modeling. An interest may need single-word term or multi-word term to represent. If we only consider single-word term, space between words can be used for segmenting interests. If we consider multi-word term, more complex term extraction algorithms and tools need to be applied.

The first question to tackle is how to get the raw-data keyword from user's message and what may be the most relevant keywords as clearly not all keywords composing a message are relevant to capture interest. The issue of keyword extraction from text is not new, therefore lots of algorithms are available nowadays that implement different strategies as explained in Chapter 3. Since our objective is to capture user interests, we are mainly interested in nouns, often represented by named entities in microposts. The other words in a text are relevant for the scoring mechanism.

After examining each online available service for keyword/named entity extraction, our choice for the tagging of microposts is AlchemyAPI [6], due to its popularity in the scientific community for the implementation of taggers and keywords extractors [Ma 2011] and the good balance it offers between accuracy and execution time. It is important to mention that AlchemyAPI may be replaced in the framework by other taggers, such as OpenCalais, which retrieves similar results.

[6] AlchemyAPI - `www.alchemyapi.com/api` - visited June 2011

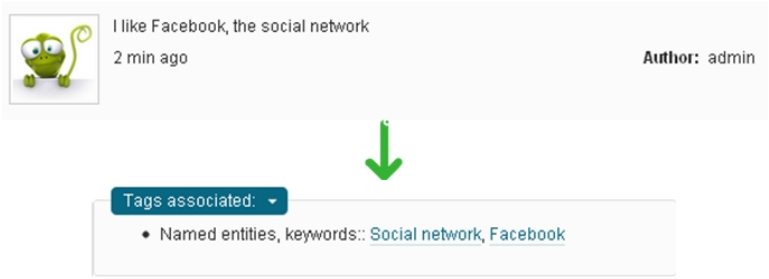


Figure 4.13: General Process to transform social content to concepts with Alchemy

AlchemyAPI is a product of Orchestr8, LLC, a provider of semantic tagging and text mining. Term extraction by AlchemyAPI is based on statistical natural language processing and machine learning. AlchemyAPI Named Entity Extractor (NEE) is one of the state-of-art tools for named entity extraction. We conducted an extensive survey of both state-of-art academic and commercial tools and found that AlchemyAPI NEE is one of the best among them and provides the best balance of extraction accuracy and processing speed.

AlchemyAPI NEE is capable of identifying people, companies, organizations, cities, geographic features, and other typed entities within HTML text, plain text, or web-based content. It supports entity disambiguation which links the extracted entity to its corresponding entry in external database including DBpedia, Freebase, OpenCyc etc. It also supports entity type identification for 34 different entity types such as automobile, city, facility.

It is however important to note that lots of other online services are available for performing keyword/entity extraction[7].

In the context of producing an input (i.e. keyword) to the algorithm, we configure Alchemy to return only keywords and entities it can find in a text. Figure 4.13 show an example of how Alchemy works. For the text "I like Facebook, the social network", Alchemy returns 2 keywords: "Facebook" and "Social network". We use such items as input data and perform its processing on a semantic level with dedicated algorithms.

Our algorithm consists of 2 sub-algorithms, one searching for the candidate concepts and one for selecting the right concepts using the contextual cues. The algorithm can be described with the following steps:

4.3.3.2 From keywords/entities to candidate concepts

The first thing to do is to search for concepts related to the keywords found. We call them the candidate concepts. For a keyword, there are always several candidate con-

[7] In the Appendix of this document we present the most interesting ones that we studied for our selection process and present their main features

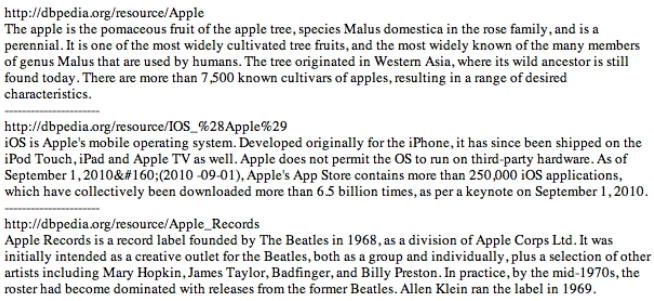

http://dbpedia.org/resource/Apple
The apple is the pomaceous fruit of the apple tree, species Malus domestica in the rose family, and is a perennial. It is one of the most widely cultivated tree fruits, and the most widely known of the many members of genus Malus that are used by humans. The tree originated in Western Asia, where its wild ancestor is still found today. There are more than 7,500 known cultivars of apples, resulting in a range of desired characteristics.
---
http://dbpedia.org/resource/IOS_%28Apple%29
iOS is Apple's mobile operating system. Developed originally for the iPhone, it has since been shipped on the iPod Touch, iPad and Apple TV as well. Apple does not permit the OS to run on third-party hardware. As of September 1, 2010 (2010 -09-01), Apple's App Store contains more than 250,000 iOS applications, which have collectively been downloaded more than 6.5 billion times, as per a keynote on September 1, 2010.
---
http://dbpedia.org/resource/Apple_Records
Apple Records is a record label founded by The Beatles in 1968, as a division of Apple Corps Ltd. It was initially intended as a creative outlet for the Beatles, both as a group and individually, plus a selection of other artists including Mary Hopkin, James Taylor, Badfinger, and Billy Preston. In practice, by the mid-1970s, the roster had become dominated with releases from the former Beatles. Allen Klein ran the label in 1969.

Figure 4.14: DBpedia Lookup service
Example of DBpedia Lookup result for "apple"

cepts in Linked Data (e.g. if keyword "apple" refers to the fruit apple, the corresponding concept will be "`http://dbpedia.org/resource/Apple`", otherwise if it refers to the company Apple, the concept will be "`http://dbpedia.org/resource/Apple_Inc.`".

However, there are also several ways to obtain this list of candidate concepts. We propose in our algorithm two ways to obtain this list:

- The first way is by using DBpedia Lookup[8]. This is a service from Dbpedia to look up DBpedia URIs by related keyword (Figure 4.14).

  There are two reasons to use DBpedia Lookup, the first one is that it searches for concepts that either the label or a resource matches, or an anchor text that was frequently used in Wikipedia to refer to a specific resource matches (e.g. the resource `http://dbpedia.org/resource/UnitedStates` can be looked up by the string "USA").

  The second interest is that the results are ranked by the number of in-links pointing from other Wikipedia pages at a result page (i.e. the popularity of the concept). The rank of a concept is very important when we do not have enough contextual cues to disambiguate a keyword. In this case selecting the most popular concept appears to be the best solution. It's because users generally have intention to use "something" their community already used (homophily effect [McPherson 2001b], meaning that interests generally propagate in a community from a user to another). As a consequence of this reaction, this "something" becomes more and more popular.

- The second way is by using the DBpedia disambiguation property. If a concept is ambiguous, the property "dbpedia-owl:wikiPageDisambiguates"[9] is present and provides its disambiguation concepts as follows:

[8] DBpedia Lookup - `http://wiki.dbpedia.org/Lookup` - visited June 2011
[9] Disambiguation property - `http://dbpedia.org/ontology/wikiPageDisambiguates` - visited June

Figure 4.15: Disambiguation property
Disambiguate concept of Facebook

The disambiguated concept is interesting when we have no contextual cues but we have some other keywords in the message related to the keyword to disambiguate. The reason is that the concepts provided by disambiguation property are usually very diverse in category. If the user talks about lots of topics/subjects, his/her context will be also diverse then the consequence in this case is that it becomes difficult for the algorithm to find the right concept corresponding to the right context. That is why it is useful when we have no existing context but other keywords in the message that we also call internal context. The internal context is unique and straight for only the message that contains the keyword to disambiguate (e.g. if in the message user talks about "Facebook" and "Social network", by using the keyword "Social network", it will be easy to choose the concept "Facebook" instead of others concepts (which have no relation to social network) in the list above).

The Concept Searching algorithm is then described in pseudo-code in Algorithm 1.

2001

Algorithm 1 *ConceptSearching*(k)

```
1: Required: k ≠ null
2: smConcept ← getSimpleMatchingConcept(k)
3: if (isExisted(smConcept)) then
4:    if (!isAmbiguous(smConcept)) then
5:       conceptsFinal[] ← smConcept
6:       RETURN conceptsFinal
7:    else
8:       conceptList[] ← getDBpediaDisambiguationConcept(smConcept)
9:       popularConceptList[] ← getDBpediaLookup(k)
10:   end if
11: else
12:   conceptList[] ← getDBpediaLookup(k)
13:   popularConceptList[] ← getDBpediaLookup(k)
14: end if
```

The idea is that from a keyword k, we construct 2 lists of concepts based on 2 ways we presented before. We need 2 lists to be used in different situations of user's context. We will clarify on that later in the *ConceptSelecting* algorithm.

To construct the lists, firstly we test if a simple matching for k gives a concept. A simple matching is simply a concatenation of the keyword and the prefix URL (i.e. "http://dbpedia.org/resource") of DBpedia (e.g. "apple" gives "`http://dbpedia.org/resource/Apple`"). If the simple matching existed, we continue to verify whether the concept given is ambiguous. If it is not ambiguous, it is the concept we are looking for, so we return this concept and stop the algorithm. However, if the concept is ambiguous, we construct the 2 lists: (i) one with the concepts coming from DBpedia disambiguation property (i.e. conceptList), (ii) another with the concepts come from DBpedia Lookup service (i.e. popularConceptList).

If the simple matching concept for k does not exist, we still construct two different lists (i.e. *conceptList* and *popularConceptList*) but the contents stay the same, those are the concepts that come from DBpedia Lookup service.

In the following section, we present the second sub-algorithm which focus on how we select concepts from the list of candidate concepts obtained from the Concept Searching algorithm.

4.3.3.3 Concept Selecting Using Cosine Similarity Computing

The core of our Semantic Matching algorithm is the Concept Selecting that employs cosine similarity - a common information retrieval similarity measure to compute the similarity between each candidate concept and the user context in order to find the

most relevant one. To do that, the concept and the user context are represented by means of vectors, which then can be compared by measuring the angle between them.

Vector Representation .

The semantic selection process's objective is to choose the most relevant concepts among all the candidate concepts that have been found for a keyword. This process is carried out taking into account the context of the user who posts the message that contains the keyword. Thus, we need to define the following model that include information about the problem as a tuple:

$$X = < U, K, C, W >$$

where U is the set of users, K is the set of keywords, C is the set of concepts, W is a set of normalized words. To clarify what normalized words could be, we now introduce here the notion of a concept abstract, as we presented before, a concept is represented by an URI and a set of properties and values. Abstract is one of these properties and is considered as the default property for every concept in DBpedia. Each concept has an abstract that resumes the essential informations of the concept. For each concept, we used its abstract as its representation, all the calculation concern a concept is then performed on its abstract. The elements of the set W is constructed by obtaining the abstract, tokenizing[10] into words, remove the stop words[11], applying the Porter Stemming algorithm[12] to normalize. The following example is the normalized abstract of the concept Facebook - a list of normalized terms:

$abstract_{Facebook}$ = {facebook, social, network, websit, launch, februari, 2004, oper, privat, own, facebook, juli, 2010, facebook, 500, million, activ, user, person, fourteen, world, user, creat, person, profil, add, user, friend, exchang, messag, includ, automat, notif, updat, profil, addition, user, join, common, interest, user, group, organ, workplace, school, colleg, characterist, servic, stem, colloqui, book, student, start, academ, year, univers, administr, intent, help, student, facebook, declar, 13, year, regist, user, websit, facebook, found, mark, zuckerberg, colleg, roommat, fellow, comput, sscienc, student, eduardo, saverin, dustin, moskovitz, chri, hugh, websit, membership, initi, limit, founder, harvard, student, expand, colleg, boston, area, ivi, leagu, stanford, univers, gradual, ad, suport, student, univers, open, school, student, final, ag, 13, facebook, met, controversi, block, intermitt, countri, includ, pakistan, syria, peopl, republ, china, vietnam, iran, north, korea, ban, place, work, discourag, employe,

[10] Tokenization is the first step in preprocessing on Information Retrieval. Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.

[11] Stop words are terms that appear so frequently in text that they lose their usefulness to be indexed as search terms.

[12] Porter Stemming algorithm official website - `http://tartarus.org/~martin/PorterStemmer` - visited June 2011

wast, time, servic, facebook, privaci, issu, safeti, user, comprom, time, facebook, settl, lawsuit, claim, sourc, code, intellectu, properti, site, involv, controversi, sale, fan, friend, januari, 2009, comper, studi, rank, facebook, social, network, worldwid, monthli, activ, user, myspac, entertain, weekli, put, end, deca, list, earth, stalk, ex, rememb, worker, birthdai, bug, friend, plai, rous, game, scrabul, facebook}

In order to compute the similarity between context and candidate concept, we define the following datasets in order to transform our datas into vectors:

- Concepts(k $\in$ K) = {c_i : $c_i \in$ C }. The set of candidate concepts for a keyword k.
- History(u $\in$ U) = {$c_{historic}$: $c_{historic} \in$ C }. The set of concepts that user used before.
- Abstract(c $\in$ C) = {w_j : $w_j \in$ W }. The set of words represent a concept c_i (e.g. example of abstract of Facebook above)
- Voc(k $\in$ K) = $\cup$ Abstract(c_i) : $c_i \in$ Concepts(k). The set of words of all the concepts corresponding to the keyword k. Voc stands for vocabulary.

In addition, we also need to define the user context. The user context is every interaction that the user had before the moment he/she make the actual interaction (i.e. post a message, reply a comment, etc.). The old interactions were at a moment like the actual interaction so they were also transformed in concepts and saved in our system. The context is therefore consisting of normalized terms as the vocabulary (i.e. $Voc(k)$) and described as follows:

- Context(u $\in$ U) = $\cup$ Abstract($c_{historic}$) : $c_{historic} \in$ Historic(u). The set of terms of all the candidate concepts corresponding to the keyword k. (Voc stands for vocabulary).

Now we can define for a user, his/her context and the concepts associated to a keyword using vectors. Those vectors are in $\Re^{|Voc(k)|}$ and each element holding a position in the vector corresponds to a term in $Voc(k)$.

- The context vector: $V_{context} = (v_i)$, where $1 \leq i \leq |Voc(k)|$ and $v_i = 1$ if the corresponding term w_i in Voc(k) appears in Context(u), otherwise $v_i = 0$.
- The concept vector: $V_{c_i} = (v_i)$, where $1 \leq i \leq |Voc(k)|$ and v_i is the frequency of the corresponding term w_i in Voc(k).

By doing this, when we want to select a concept from the candidate list, we can create a $V_{context}$ for the context of the user and then a $V_{concept}$ for each concept related

to the keyword. Then, we can compare $V_{context}$ with each $V_{concept}$ using a cosine similarity measure. The cosine of the angle between two vectors is a value between 0 and 1. When the angle is small the cosine value tends to 1, when the angle is big the cosine value tends to 0. Smaller the angle is, more relevant the concept will be. The formula to compute the similarity is the standard cosine similarity:

$$Sim(V_{context}, V_{concept}) = cos\theta = \frac{\overrightarrow{V_{context}}.\overrightarrow{V_{concept}}}{|\overrightarrow{V_{context}}|.|\overrightarrow{V_{concept}}|}$$

Concept selection algorithm We now know how to build a context vector and concepts vector. We also know how to compare between them to select the most relevant concepts. The Concept Selecting algorithm is described in Algorithm 2.

The principal situations that may occur while executing the algorithm are the following:

- **user context does not exist**
  - **if other keywords exist** we consider these keywords as context, construct the context vector, and compare it with the concepts from DBpedia disambiguation property. We return to the question of having 2 lists of concepts. The list of concepts that comes from DBpedia disambiguation property is used only when we have no user's context but some other keywords in the same message. By computing the similarity we have a score for each concept (line 10). We continue to look at the scores:
    - **other keywords have no influence** it's when all the score is equal to 0 meaning that those other keywords in the message are useless, they have no relation neither with the keyword we are working on. In this case the most popular concept from DBpedia Lookup service will be returned as result (line 17).
    - **other keywords give scores** in this case, we return the concepts with highest scores (line 19).
  - **no other keywords** in this case the most popular concept from DBpedia Lookup service will be returned as result (line 22).
- **The user context exists** we then add other keywords (if they also exist) to the context and then construct the context vector. For each concept from DBpedia Lookup, construct a concept vector and compute the similarity. Same as before, we compare the scores and return the concepts with the most highest scores (line 32).

Algorithm 2 *ConceptSelecting*(*user*, *conceptList*, *popularConceptList*)

```
1: Required: conceptList ≠ null & popularConceptList ≠ null
2: voc[] ← constructVocabulary(conceptList)
3: context[] ← getContext(user)
4: otherKeywords[] ← getOtherKeywords()
5: if (isEmpty(context[])) then
6:    if (!isEmpty(otherKeywords[])) then
7:       contextVector[] ← constructContextVector(otherKeywords[], voc[])
8:       maxScore ← 0
9:       for all (concept in conceptList[]) do
10:         conceptVector[] ← constructConceptVector(voc[])
11:         sim ← getSimilarity(contextVector[], conceptVector[])
12:         conceptAndScoreList ← (concept, sim)
13:         if (sim ≥ maxScore) then
14:            maxScore ← sim
15:         end if
16:      end for
17:      if (maxScore = 0) then
18:         RETURN popularConceptList[0]
19:      else
20:         RETURN getBestConcept(conceptAndScoreList)
21:      end if
22:   else
23:      RETURN popularConceptList[0]
24:   end if
25: else
26:   context[] ← add(otherKeywords[])
27:   contextVector[] ← constructContextVector(context[], voc[])
28:   for all (concept in popularConceptList[]) do
29:      conceptVector[] ← constructConceptVector(voc[])
30:      sim ← getSimilarity(contextVector[], conceptVector[])
31:      conceptWinnerList ← (concept, sim)
32:   end for
33:   RETURN getBestConcept(conceptAndScoreList)
34: end if
```

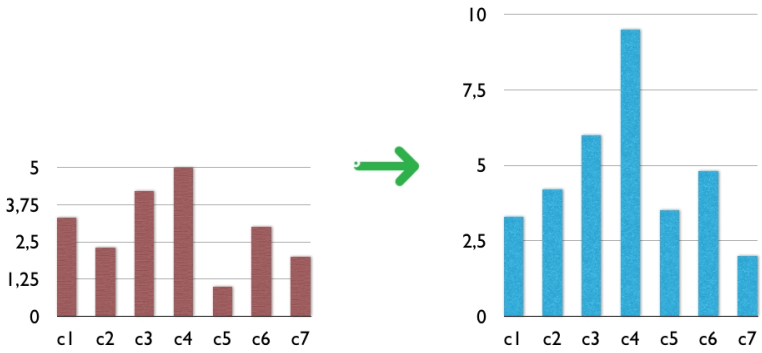


Figure 4.16: Illustration of the winner concepts.

One thing to be noticed here is that we don't return only the concept with the highest score but n concepts with n highest scores. Therefore, one or several concepts will be returned based on the difference of their score with the others. By default, we designed our algorithm to return concepts with scores higher than 80% of the highest score to be sure that we won't miss any concept. The reason behind this idea can be described in Figure 4.16.

The figure in the left shows the usual case of similarity scores when there is not many context in the system for a user when processing his/her message. The scores are therefore always low and there is not a big difference between those scores. We then select not only the ones with the highest scores (i.e. the concept c4) but the 2 concepts c3 and c4 whose scores are larger than 80% of the score of c4. We update the context with these new concepts. The user context then continues to grow in this way and becomes more and more stable and oriented to a few principal subjects/topic. Once the context is rich, the distinction between the similarity scores will appear and can be described as shows the right hand side of the Figure 4.16.

In the following section, we present the second step of the profile construction process. Once the user's interactions (i.e. update, post, message, etc.) are matched to DBpedia concepts, we expand these concepts to better approach user's interests.

4.3.4 Semantic Expansion in the Knowledge Base

4.3.4.1 Context

The main role of this operation is to allow us better categorizing the users' profiles and to better approximate their interests. As an example, a user who shared about topics such as Facebook and Twitter might have a general interest in Web 2.0 technologies, which can be inferred by propagating these atomic interests to more general categories. In addition, semantic expansion is useful for the enrichment of a concept with concepts that are related and provide additional context to the user's interests and expertise.

4.3.4.2 SemEx (Semantic Expansion) Algorithm

The first step in this operation consists in building a "semantic sphere" associated to the explicitly shared concept (Figure 4.17), that contains all candidate concepts that will form the expansion. In our approach, we explore three types of connections in Linked Data to construct this kind of sphere:

- the first and the most interesting is represented by hierarchical links to category concepts (e.g. concept "Gran Torino" will have "Gang films" and "American drama films" as hierarchical expansions). They generalize a concept to categories, e.g. someone interested in the film "Gran Torino" may also be interested in other "Gang films".
- the second dimension is that we explore concepts connected to each category concept that was previously retrieved (e.g. the movie "Punisher", the neighbor of "Gran Torino" in the "Gang films category").
- the third dimension explores concepts directly connected to the initial concept in the knowledge base. These concepts come from the Infobox properties[13] of the concept. The wikipedia infobox contains the most relevant and specific information about a wikipedia article. By expanding in this dimension we got the closest concepts to the initial concept (e.g. "Clint Eastwood" is the director and main actor of movie "Gran Torino").

By these expansions, with a given concept we retrieve the concepts from its generality, the concepts of the same category and also those which are relevant and specific to the concept.

The formal definition of the semantic expansion of a given concept c in a knowledge base $\mathcal{K}$ is the following:

$$E_{\mathcal{K}}(c) = \{\forall c_{exp} \in \mathcal{K}, \exists p \in P_{\mathcal{K}} | p(c, c_{exp})\}$$

Based on the knowledge base $\mathcal{K}$, different properties are used to expand the concept. A property p belongs to the set of properties $P_{\mathcal{K}}$ of a concept c which is defined as follows:

$$P_{\|} = \{subject(c)^{14}, property(c)^{15}, isbroaderof(subject(c))^{16}\}$$

We have now all the necessary elements to expand a concept and the Semantic Expansion algorithm is shown in Algorithm 3.

[13] Properties mapping from Wikipedia to DBpedia knowledge base.
[14] `http://purl.org/dc/terms/subject`
[15] `http://dbpedia.org/property`
[16] `http://www.w3.org/2004/02/skos/core#broader`

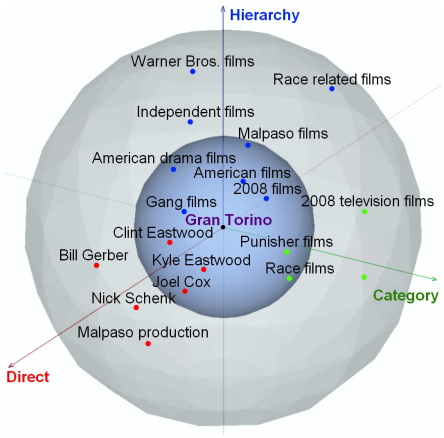


Figure 4.17: Semantic Expansion sphere
Dimensions of the semantic expansion of the concept "Gran Torino" - a movie directed by Clint Eastwood in 2010.

Algorithm 3 $SemanticExpansion(c, k)$

Required: $c \neq null$
$conceptList[] \leftarrow []$
$hierarchyConcepts[] \leftarrow getHierarchyConcept(c)$
$categoryConcepts[] \leftarrow getCategoryConcept(c)$
$directConcepts[] \leftarrow getDirectConcept(c)$
$conceptList[] \leftarrow add(hierarchyConcept, categoryConcept, directConcept)$
for all ($concept$ in $conceptList[]$) **do**
    $sim \leftarrow getSimilarity(c, concept)$
    $conceptAndScoreList[] \leftarrow add(concept, sim)$
end for
RETURN $getBestConcept(conceptAndScoreList[], k)$

We first expand the concept c in the 3 dimensions by using SPARQL to query DBpedia with the properties that we mentioned before. An example SPARQL query that allows to retrieve all concepts (?category) that are connected to a given concept (e.g. Facebook) with a given property (e.g. purl:subject) is as follows:

PREFIX purl: <http://purl.org/dc/terms/
SELECT ?p ?category
WHERE
<http://dbpedia.org/resource/Facebook> purl:subject ?category

Figure 4.18: Hierarchy Expansion for Facebook concept
All the concepts in the hierarchy dimension.

Figure 4.19: Category Expansion for Facebook concept
All the concepts in the category Web 2.0 of Facebook.

For example, with the Facebook[17] concept, the expansion in the hierarchy dimension (i.e. SPARQL query with *subject* property) gives the result shown in (Figure 4.18).

For a category found in the first expansion, we are looking for the concepts belonging to that category. The expansion in the category dimension (i.e. SPARQL query with *isbroaderof* property) gives the result shown in Figure 4.19.

Finally, we expand the concept from the Facebook infobox. Figure 4.20 show the real infobox of Facebook in wikipedia, the concepts from this infobox are: Mark Zuckerberg[18], Dustin Moskovitz[19], Chris Hughes[20] (founders of Facebook).

The concepts received are put in the *concepList*. However, a knowledge base often contains a very large amount of data from the very specific to the very general. Therefore the result set obtained from a semantic expansions operation is always large (represented by the external sphere in Figure 4.17.

In our approach we want to keep only a subset of concepts that are the closest ones to the user's interests (inner sphere in Figure 4.17). To perform the filtering, we compute the similarity of each concepts present in the expansion with the the concepts

[17] http://dbpedia.org/resource/Facebook
[18] `http://dbpedia.org/page/Mark_Zuckerberg`
[19] `http://dbpedia.org/page/Dustin_Moskovitz`
[20] `http://dbpedia.org/page/Chris_Hughes_%28entrepreneur%29`

Figure 4.20: Direct Expansion for Facebook concept
Facebook infobox

associated to the user's social update (Algorithm 3 - line 7). The calculation is exactly the same with those we did in the Semantic Matching algorithm where the similarity between the concepts is computed by the cosine similarity between their abstracts [21]. After having each concept associated to a score which represents its relevancy to the initial concept, we return the $top - k$ concepts having the best scores.

The fact that we compute the similarity score with the abstracts is based on the idea that an abstract will normally contain lots of keywords related to the concept. These keywords build up a small vocabulary which can serve as a local context in order to find the closest concepts from the expansion (e.g. The abstract of the film Gran Torino contains essentially keywords like "American", "drama" and more frequently "Eastwood"). When compared to this local vocabulary, the concepts like "American drama" or "Clint Eastwood" will certainly score more points in similarity than the others like "Fictional American people of Polish descent", also present in the expansion set. Once we have the similarity scores, we rank the concepts by sorting them and add the concepts with the $top - k$ highest scores to user's profile.

With this heuristics performed on the expansion set, we can significantly reduce the amount of concepts that might be irrelevant for the user in terms of expertise.

In the following, we present the second algorithm strongly related to semantic expansion. The previous algorithm provides us with a set of concepts from the semantic neighborhood of a concept and that are most relevant for the user. Suppose each concept in the user profile has an associated weight. The question is how to propagate those weights to these new concepts?

[21] An abstract of a concept in DBPedia is equal to its textual definition

4.3.5 Propagation of Concept Scores with a Constraint Spreading Activation Algorithm

In the previous sections, we presented the general architecture and the main steps for concept extraction from microposts:(i) Semantic Matching (i.e. matching keywords and named entities in the microposts to semantic concepts in a Linked Data knowledge base) and (ii) Semantic Expansion (i.e. the extension of the initial set of concepts with additional concepts, exploring the relations in said Linked Data knowledge base).

The result of the semantic expansion of user profile concepts is the fact that new concepts are introduced in the profile. Said concepts were not explicitly shared by the user, so we have no information about corresponding sentiment, entropy or term frequency. In other words, we need to approximate the level of expertise and interactivity of the user with regards to these concepts in the expansion set. In order to achieve this, we consider existing approaches in the area of explicit semantic preference spreading algorithms, introduced in Section 3.4.

Our approach is largely inspired by the early works on Constrained Spreading Activation [Cohen 1987] [Crestani 2000]. We adapt these algorithms to the specific context of Linked Data and introduce rules for the preference activation that take into account the names of properties that link concepts in the knowledge base.

The process is as follows: (i) the SemEx algorithm identifies the most relevant semantic neighborhood of a given concept and (ii) the spreading activation algorithm propagates the right scores to be associated to said concepts, depending on the type and number of relationships that exists between them.

To illustrate the challenge, consider concepts "Twitter" and "Facebook" in the user profile. Both concepts are connected to concept "Web 2.0", that is not directly in the profile, only in the expansion set of both said previous concepts.

Consider the expertise of the user in concept Twitter 0.6 and in concept Facebook 0.8. What should be the expertise of the user in concept Web 2.0?

The activation of concepts in the expansion set that could be relevant for the user is based on an approximation to conditional probabilities. The formal definitions necessary to the algorithm are as follows:

- Let $p_u(c_x) = u_x \in \mathcal{D}$, domain of scores associated to user profile concepts, i.e. $[0, 1]$. u represents the user, $u \in \mathcal{U}$ and c a concept in the knowledge base $\mathcal{K}$.
- The probability that c_x is relevant for the user can be expressed in terms of the probability that c_x and each concept c_y directly related to c_x in the knowledge base belong to the same topic, and the probability that c_y is relevant to the user.
- With this definition, the relevance of c_x for the user can be computed by a standard CSA algorithm, starting with the initial set of semantic concepts P_u in the user profile, i.e.,

$$P_u = \{c_k \in \mathcal{O} | p_u(c_k \neq 0\}.$$

- Let $\mathcal{R}$ be the set of all relations in $\mathcal{K}$. The spreading strategy is based on weighting each semantic relation $r \in \mathcal{R}$ with a measure $w(r, c_x, c_y)$ that represents the probability that given the fact that $r(c_x, c_y)$ holds, c_x and c_y belong to the same topic.
- This is used for estimating the relevance of c_y when c_x is relevant for the user.
- The weight $w(r, c_x, c_y)$ is interpreted as the probability that c_y is relevant for the user if we know that the concept c_x is relevant for the user, and $r(c_x, c_y)$ holds.
- With this measure, concepts are expanded through the semantic relations of the knowledge base, using a constrained spreading activation mechanism over the semantic network defined by these relations.
- As a result, the initial set of concepts P_u is extended to a larger vector EP_u, having $EP_u[c_k] \geq P_u[c_k]$ for all $c_k \in \mathcal{O}$.
- Let $\mathcal{R}^{-1}$ be the set of all inverse relations of $\mathcal{R}$, i.e., a concept c_x has an inverse relation

  $$r^{-1}(c_x, c_y) \Leftrightarrow \exists r(c_y, c_x) | r \in \mathcal{R}.$$

  Let

  $$\hat{\mathcal{R}} = \mathcal{R} \bigcup \mathcal{R}^{-1} = \mathcal{R} \bigcup \{r^{-1} | r \in \mathcal{R}\} \text{ and } w : \mathcal{R} \rightarrow [0, 1].$$

- The extended concept vector EP_u is computed according to two possible situations:
  - $EP_u(c_y) = P_u(c_y)$, if $P_u(c_y) > O$.
  - $R(\{EP_u[(c_x)] * w(r, c_x, c_y) * power(c_x)\})$, otherwise
- where $power(c_x)$ is a propagation power assigned to each concept c_x (by default equals to 1) and
- $R(X) = \sum \{(-1)^{(|S|+1)} * \prod(x_i)\}$

The following example shows its simplicity of practical use. Suppose, the user has shared about two of these concepts, which are related to a third through two different relations in the knowledge base. The expansion shows how a third value is inferred, accumulating the evidence of relevance from the original two preferences.

- We consider concept c_x = "Facebook" and concept c_y = "Twitter".
- The concepts c_x and c_y are both expanded to concept c_z = "Web 2.0".
- The inferred expertise value of the user for concept c_x is 0.6 and concept $c_y = 0.4$.
- The objective is to propagate an expertise value to concept c_z, having the expertise values for concepts c_x and c_y.
- Consider the relations between concepts c_x and c_z of being both "<http://purl.org/dc/terms/subject>".
- We attribute the propagation decay for the relation "<http://purl.org/dc/terms/subject>" of 0.6 (i.e. we consider generalization as an important factor of propagating expertise).

According to these criteria, concept
c_z = "Web 2.0" will be attributed the following expertise value:
First we compute

$$p_z^1 = p_z * w(r_1) \text{ and then } p_z = p_z^2 = p_z^1 + (1 - p_z^1) * p_y * w(r_2).$$

Inserting the corresponding values results in:

$$p_z^1 = 0.6 * 0.5 = 0.3 \text{ and}$$
$$p_z = p_z^2 = 0.3 + (1 - 0.3) * 0.4 * 0.5 = 0.3 + 0.7 * 0.4 * 0.6 = 0.468.$$

In the following, we describe the set of parameters that have been included in the algorithm in order to avoid cases of excessive semantic propagation.

- ε - Minimum Threshold Weight. First of all, a given concept must have a minimum threshold weight in order to expand its weight to related concepts. This is important, as a high threshold value improves the performance of the spreading algorithm (e.g. few concepts to expand). However, very high values result in the fact that the underlying semantics of the knoweldge base will be not explored, resulting in poorer propagation inferences.
- n_e - The maximum number of expansion steps to be performed by the spreading algorithm.
- n_h - The maximum number of times a concept can be generalized. This parameter is equivalent to ne applied to hierarchical relations, like subject in DBPedia. Once a concept has been expanded up to n_h hierarchical levels, it would be convenient not to expand it more. The intention of this constraint is to not generalize an expertise (semantically) too much, as this type of expansion is a risky assumption with the original user's expertise.

- n_f - The maximum fan-out (i.e., number of output properties) a concept can have to be expanded. The aim of this constrain is to reduce the "hub effect" in concepts with many relations to other concepts.
- $power(c_x)$ - The propagation intensity (strength) of a concept. This factor multiples the effect of propagating the concept weight. By default, it is set to 1.
- $w(r, c_x, c_y)$ - The propagation decay of a relation between two given concepts. This parameter approximates the probability that a concept c_y is relevant given that c_x is relevant and relation $r(x, y)$ holds. It can be seen as the propagation power of the relation $x \in \mathcal{R}$ for concepts c_x and c_y.

Having these definitions, the algorithm is as follows:

Algorithm 4 *SemanticPropagation*(P, EP, w)

```
1: //init the expanded concept weights with the input ones
2: for all (concept in conceptList[P]) do
3:    EP[c_x] = P[c_x]
4:    //create a priority queue based on concept weights (intially null)
5:    Q ← buildPriorityQueue(O*{prev = 0, hierarchylevel = 0, expansionLevel = 0}
6:    while (Q.isEmpty == false)
7:    //Identify the next concept to expand
8:    (c_x, prev_x, hierarchylevel, expansionlevel) ← Q.pop()
9:    //Check the minimum concept weight constraint
10:   if (EP[c_x] < ε) exit
11:   check the maximum expansion constraint
12:   if(expansionLevel ≥ n_e) GOTO while
13:   /*retrieve the neighbourhood of the current concept from the
14:   database (constructed by the SemEx algorithm presented before)*/
15:   for all (r, c_y inc_x.Neighborhood do
16:      check the hierarchical level expansion constraint
17:      if EP[c_y = 1 OR (r.isHierarchical() AND hierarchyLevel ≥ n_h))
18:      GOTO for
19:      //undo the last update from c_x
20:      EP[c_y] ← (EP[c_y] − w(r, c_x, c_y) * power(c_x) * prev_x) /
21:      (1 − EP[c_y] * w(r, c_x, c_y) * w_f(c_x, n_f) * prev_x
22:      //recompute the propagation score value for the concept
23:      EP[c_y] ← (EP[c_y] + (1 − EP[c_y]) * w(r, c_x, c_y))*w_f(c_x,n_f)*power(c_x)*EP[c_x]
24:      if (r.isHierarchical()) hierarchyLevel++;
25:      Q.push(c_y, prev_y, hierarchyLevel, expansionLevel)
26:   end for
27:   expansionLevel++
28: end for
```

Finally, we highlight the fact that the approach is semi-automatic, users having the option to annotate their profile and thus, approve or disapprove concepts. If a concept is not validated by the user, it is inserted into a special table of the database and the given concept and corresponding fragment of Linked Data will not be explored any more.

4.4 Concept Scoring Mechanism

As mentioned before, our scoring mechanism's objective is to capture both interactivity and expertise. In order to achieve this, we consider different measures on a microposts

that we describe in the following sections. The fact that we employ these scores is the result of several experimentation of microposts that show that there are the most interesting measures and these distinguish the best experts.

It is important to mention in a first place that these individual scores can be divided into two categories:

- *Statistical Measures.* In this category of scores are included standard statistical measures used in information retrieval and document analysis. More specifically this consists of the traditional tf-idf score.
- *Semantic Measures.* In this category of scores are included semantic measures on the style of a post in a given domain. This consits of (i) sentiment polarity analysis and (ii) entropy analysis.

The different scoring mechanisms are available individually in the framework and can be manually activated or deactivated according to the needs of a specific applications that is constructed on top.

In the following, each individual score and its role will be shortly depicted.

4.4.1 Term Frequency / Inverse Document Frequency Score - $TF - IDF$

This weighting score identifies (i) concepts that better describe user u and (ii) those concepts which better distinguish him from the other users.

First we measure the frequency of each concept c for a user u. We call this factor as *Item Frequency - IF*. IF represents the number of times item c occurs in the profile of user u divided by the total number of items in the user profile. The User Frequency (UF) is the number of users in which concept c occurs at least once. Finally, the Inverse User Frequency $IUF(c)$ can be calculated from $UF(c)$ as follows ($|U|$ is the total number of users):

$$IUF_c = log\frac{|U|}{UF_{(c)}}$$

According to this equation, the IUF is low if the concept occurs in many user profiles, whereas it is high if the concept occurs in few profiles.

Finally, the new weighted value of concept c in the profile of user u is calculated as following:

$$W_{u,c} = IF_{u,c} * IUF_c$$

Figures (4.21 4.22) show the frequency of sharing in a given subject area for a given user. We can observe the difference in sharing patterns between a set of concepts.

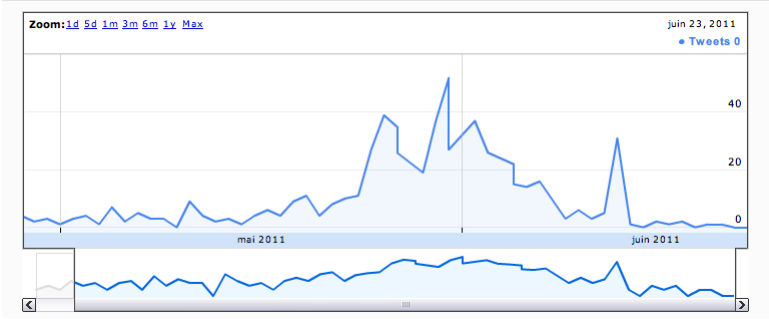


Figure 4.21: Daily Sharing Patterns of a user related to a given concept, i.e. Facebook

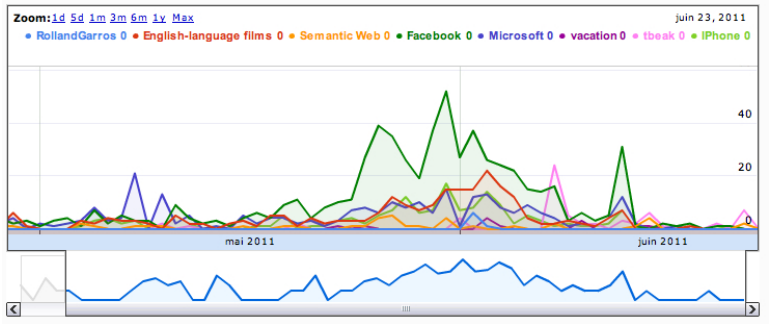


Figure 4.22: Comparison of daily sharing patterns between the most popular and frequent concepts shared by the user. Each color represents the sharing pattern for a given topic.

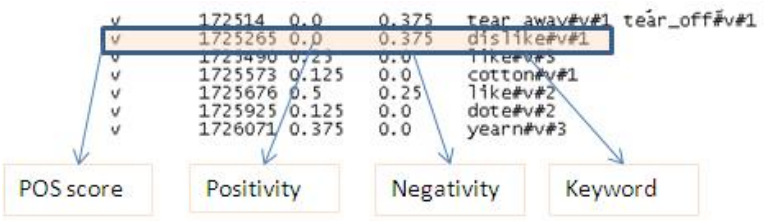


Figure 4.23: Example of structure of the Sentiwordnet Database

4.4.2 Sentiment Polarity Analysis - S

In this section we introduce the first component of our profile scoring mechanism, the sentiment polarity.

Sentiment polarity analysis consists in extracting the sentiment score associated to a message. This is based on keywords that express subjective feelings that can be generally found in the neighbordhood of named entities in a micropost (e.g. like, hate, bad, good, nice, beautiful etc.). In our case, we introduce the sentiment polarity in the weighting schema, as an interaction expressing strong sentiment about an object may reflect higher motivation for the user to engage in an interaction, as the said object may have had a high impact on the user's state of mind.

An interesting resource for sentiment analysis is the SentiWordNet vocabulary [Esuli 2006b]. SentiWordNet is a lexical resource for opinion mining which is publicly available. SentiWordNet assigns to each synset of WordNet two sentiment scores: positivity and negativity (Figure 5.5). This is a numerical value in the interval [0,1], indicating how positive or negative polarity a given keyword represents.

The main difficulty in using such a dictionary, such as Wordnet is the fact that a given term may appear in different synsets. Therefore, the right synset needs to be first selected before retrieving the sentiment polarity of the word. Part-of-Speech (POS) [Sun 2011] tagging is used to annotate the grammatical category of words in the message using the Stanford POS tagger [22]. Based on such grammatical patterns, we select the synset from the vocabulary, which has the same grammatical category associated. In such way, we can distinguish between the meaning of the verb "like", expressing a strong positive sentiment and the adjective "like", used as a comparison (the shape of the apple is like the orange), but with no value in terms of contribution to the sentiment polarity of the micropost, as it is a neutral, objective word.

Results with this approach are satisfactory because of the small size of the vocabulary used to express sentiments in microposts. For computing the sentiment of a message, we consider the average of the individual sentiment values that are associated to each synset in the SentiWordNet vocabulary.

[22]Stanford POS tagger - `http://nlp.stanford.edu/software/tagger.shtml` - visited April 2010

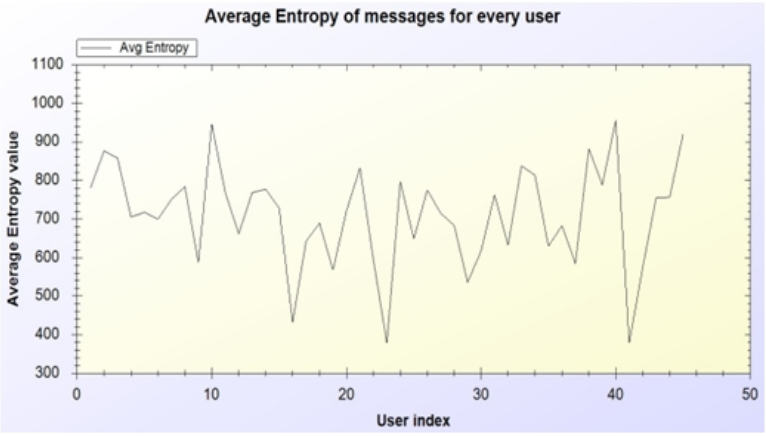


Figure 4.24: Example of variations of the entropy of posts related to a given concept. Axis X represents the users, axis Y the corresponding average entropy of posts of the user related to the given concept, i.e. "company".

4.4.3 Entropy Analysis of Microposts - E

In this section we introduce the second component of our scoring mechanism, called entropy. In this score, we consider the question of how rich is the vocabulary employed by the user to talk about a specific subject area? Entropy seems a promising measure to compute the statistical complexity of a post, by taking into account the entropy of words in the post. The following formula (inspired from information theory [Shannon 1948]) shows that for a post p_j with λ number of words what is the entropy of p when each word has frequency p_i:

$$entropy(c_{pj}) = 1/\lambda * \sum p_i * [log(\lambda) - log(p_i)]$$

Figures 4.24 and 4.25 show the variation of entropy for a user community and for two specific concepts, i.e. "'company" and "'google". These figures show the fact that people share about the same concept with different complexities. Our intuition is thus that this measure can be interesting for the identification of expertise.

It is to be noted that our computation of entropy is none-limitative and other dimensions could also be included in the computation of complexity, such as the number of named entities in a post, presence of hyperlinks, hashtags etc., which all reflect the user's motivation to share interesting content.

4.4.4 Expertise/Interactivity Score

Expertise in a given topic is composed of frequency of sharing content in the given topic, sentiment and the complexity of message shared in the given topic by the user. More

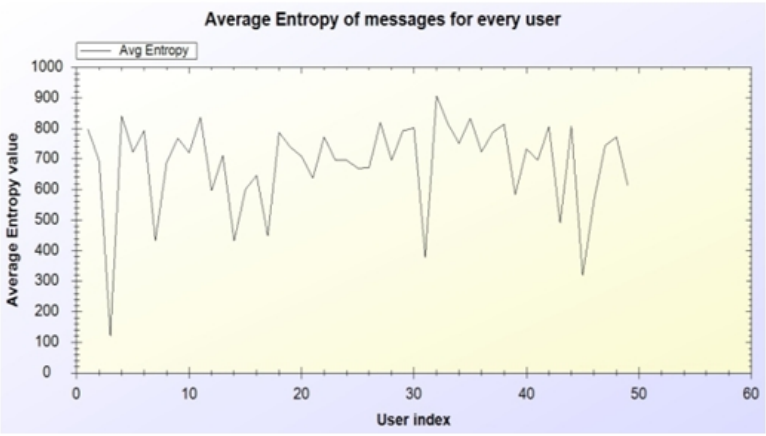


Figure 4.25: Example of variations of the entropy of posts related to a given concept. Axis X represents the users, axis Y the corresponding average entropy of posts of the user related to the given concept, i.e. "google".

concretely, in our definition of expertise, interactivity plays also a small role, measured mainly by the sentiment component. In this way, our objective is to identify first users who share frequently, who express an interesting sentiment with regards to a topic of a query and also those who share more complex and rich messages.

In order to have a global expertise score, the previously mentioned individual expertise scores must be combined. For this, a standard weighting schema may be employed that has the following shape.

First, we consider w_x a set of weights that represent the importance of a given score in the global score. Each $w_x \in \Re$. A further criteria can be introduced here that takes into account the amount of data processed for a given score. In other words, since some individual score calculations can take a lengthy time to process (e.g. the computation of sentiment values), they can be estimated or omitted at a given time t, depending on how much data has been processed with regards to the total amount of crawled data. The variable p_x is used to determine the progress of calculating a given score $score_x$: if p_x is 0.0, the calculation hasn't started, and p_x is not taken into account while calculating the overall score. If p_x is 1.0, the $score_x$ calculation is complete (e.g. all available interactions have been processed to compute it, so we can have maximum confidence in the result). Any value between 0.0 and 1.0 indicates that the calculation is in progress and $score_x$ is an estimate (converging on the final value).

Taking this into account, the global expertise score of a user u related to a concept c has the following shape:

$Expertise_{uc} = \frac{\sum w_x * [p_x(u)] * score_x(u)}{\sum w_x * [p_x(u)]}$

We consider that each progress score p_u and each individual score $score_u$ are in the

interval [0,1].

The confidence associated to the overall expertise of a user related to a profile concept can be computed as the overall progress of the corresponding computations:

$confidence_{uc} = \frac{\sum w_x * [p_x(u)]}{\sum w_x}$

An example weighting schema may be to give weight values from 1 to 3. In this case, the tf-idf could be of weight 3 and the others of weight 1 [23].

4.4.5 Ranking Mechanism

Once the different algorithms for the construction of user profiles from micropost data and the scoring of said profile concepts identified, the question is how to select the most relevant $top-k$ users for a query (i.e. question in natural language)?

In order to represent and manipulate user profile vectors, the most appropriate conceptual models appears to be the well-known vector-space model from the field of information retrieval [Salton 1975]. The fact that we deal with concept-based profiles implies that such vectors are composed of URI-s and not the textual form of concepts. The corresponding textual form is retrieved from the database when the tag cloud is constructed for visualization.

In this model, a set of terms representing a document d is formalized in the following way:

$$v(d) = [w_{t_1}, w_{t_2}, ..., w_{t_i}, ..., w_{t_n}],\ n \in \mathcal{N} \text{ with } w_{t_i} \in \mathcal{R}$$

In addition, an incoming query is also represented as a vector, having

$$v(q) = [w_{t_1}, w_{t_2}, ..., w_{t_i}, ..., w_{t_n}],\ n \in \mathcal{N} \text{ with } w_{t_i} \in \mathcal{R}$$

In order to enrich the query with additional semantics, the semantic matching and expansion algorithms are executed also on the query terms. For this, the vocabulary of the user asking the query is used for disambiguation, if available [24], as users should also have the possibility to ask questions without registering to the platform.

The similarity function between normalized vectors, like in traditional vector-based models, relies in our case also on cosine similarity. Thus, the relevance of a vector R (representing a user profile) with another vector Q (representing the query) is computed by:

$$Sim(V_{query}, V_{profile}) = cos\theta = \frac{\overrightarrow{V_{query}}.\overrightarrow{V_{profile}}}{|\overrightarrow{V_{query}}|.|\overrightarrow{V_{profile}}|}$$

[23] The current implementation of the system uses a weight 3 for tf-idf and 1 for entropy and sentiment

[24] the vocabulary belongs to the user profile newcomers and unregistreted users may not have this vocabulary available as they do not have any messages processed yet (see section 4.3.3.1)

4.5 Information Granularity for Privacy Management

After the presentation of the main algorithms related to the user profile construction from microposts, we adress in this section privacy in social platfroms and propose a novel generic approach that introduces the concept of granular refinements of a profile attribute depending on a bluring criteria.

In the following section we focus on the theoretical aspects of the approach and on an illustration of the concept with a specific bluring criteria.

[Barker 2009] define the privacy as a tuple $P =< p, v, g, r >$ where p stands to the purpose of the collection, v to the visibility, g to the granularity and r to the retention. Our work in this area concerns the granularity dimension and we investigate a way to exploit semantic technologies to reinforce the privacy.

The new privacy management method that we propose allows controlling the detail level of exposed information. The meaning of information details is seen as the precision or the granularity of the information that a person wants to share. As an example, the set of tags *{Alice, My girlfriend, Alice' birthday party}* is more precise than *{Alice, Friend, Party Event}*.

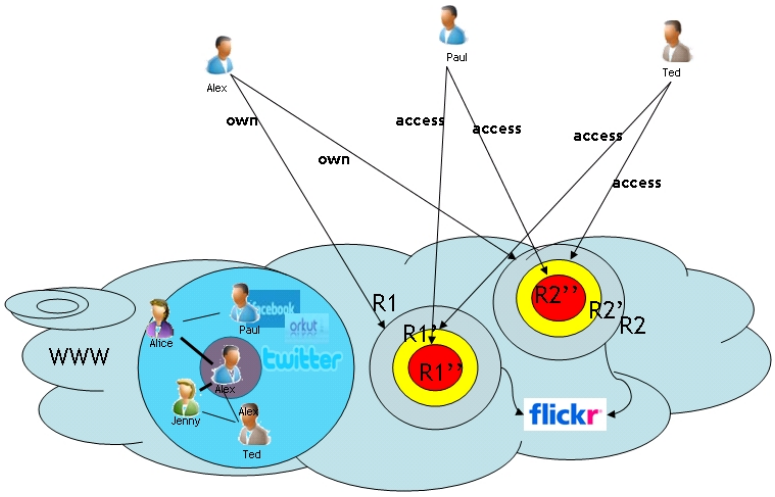


Figure 4.26: General principle of the proposed approach in social platforms

In the following, we will present our approach by considering the social relation that exists between the owner of a tag and the viewer, as this is the simplest to understand.

Figure 4.26 summarizes the proposed approach by illustrating the general view. To each tag, there is a corresponding detail level of the information that will be consulted by the tagged person(s). Let's consider the tags of a photo in a social network. A user, Alex, owns two resources, say tagged photos, and wishes to share them with two of his friends Alice and Ted as illustrated in Figure 4.26. According to the social relations

that Alex has with his two friends, Alice seems to be closer to Alex than Ted. Thus, Alex would like to expose more details about the shared resources to Paul and fewer details to Ted.

In the Figure, R_1 and R_2 are the complete resources where $R_1^{'}$, $R_2^{'}$, $R_1^{''}$, and $R_2^{''}$ are less detailed versions of the complete information. It should be noted here that $R_i^{'}$ and $R_i^{''}$ may be equal in terms of detail level to R_i if the user has a very important social relation with the different relatives.

More generally, there are three main elements composing our approach to manage privacy in social platforms: (i) Users (U)and (ii) Tags (T). These are similar to the main pillars of social platforms.

We recall in the following these three components as well as their roles in the privacy preservation.

4.5.1 The Role of Users (U) in the Granular Approach

The primary element of our proposal is the user. To a social platform correspond a set $U = \{u_1, u_2, ..., u_n\}$ of n users. In a social platform, users are (or could be) be modeled as a social graph to represent the (social) relations that may exist between them. The role of social networks is to qualify the relationship between the owner of the photo and the viewer of the photo for example. The social network of the user can be a social networking site (e.g. Facebook) or a different social networking platform (e.g. Flickr, Del.icio.us). That is, for each user we associate a social network composed of all the persons who have a direct relation with her. User's social relations are defined as a finite set of categories C. For instance, we consider $C = \{null, Family, Friends, Colleagues, Strangers\}$ for the sake of simplicity. Information regarding the category of the social relationship between the owner and the viewer can be extracted from the different social networks (e.g. Facebook already proposes this feature). Thus, we define a function $\mathcal{N}$: $U \times U \rightarrow C$ which associates for a pair of users, the category of their social relation. Note that the value *null* means that there is no relation between two nodes. Since it is well established that social relations are not symmetric (i.e. the regard you hold on others is not necessarily the regard they hold for you.), it results that the function $\mathcal{N}$ is also not symmetric. This means that, e.g. $\mathcal{N}(u_i, u_j) = Friends$ whereas $\mathcal{N}(u_j, u_i) = Colleagues$.

From the privacy point of view, social categories define a hierarchy that gives access to different information according to each social category. Thus, we define the concept of *social relations pyramid* as follows:

Social Relations Pyramid (SRP): A social relations pyramid is a structure that defines an ordering according to a social proximity in the social relations a person, or a group of persons, has with other persons or group of persons in a social environment.

The concept of SRP translates thus two aspects: (i) the fact that generally socially

people give a higher privacy degree to view people compared to all the people they have in their social network, and (ii) the different degrees of privacy that a user tend to maintain according to the different categories. Figure 4.27 shows an illustration of the *SRP* concept.

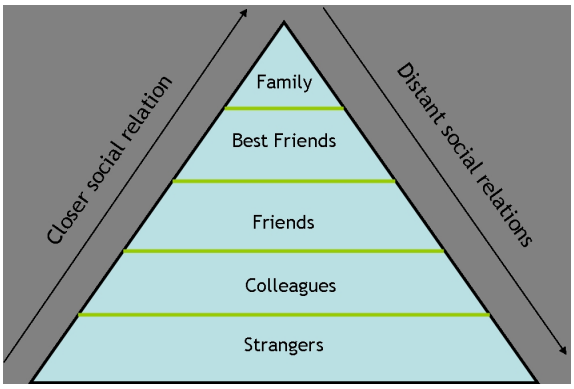


Figure 4.27: Illustration of the social relations pyramid (SRP)

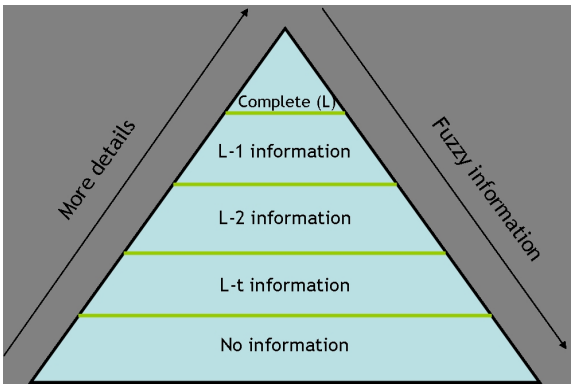


Figure 4.28: Illustration of the social privacy pyramid (SRP)

Also, it is very well known and very frequent in social networks environments that people give more interest for a specific category of people according to a specific context. For example, it is more likely that users share information about personal life with their best friends rather than with their colleagues. Inversely, it is more likely that that a person shares professional information with their colleagues rather than their friends or family. In other words, this adds a dynamic dimension to the the modeling of the social relations of users which is controlled by the current context of the user.

To capture this situation, we define the concept of *dynamic social relations pyramid (DSRP)*.

Dynamic Social Relations Pyramid (DSRP)) : A *DSRP* is a *SRP* which is conditioned by an additional parameter which makes it able to calculate and update the configuration of a *SRP* according to, e.g., the (social) context of the user.

4.5.2 The Role of Tags (T) in the Granular Approach

A tag is a keyword which participates in the explanation of the meaning of a resource. A tag is generally associated by a user to resource.

To introduce a privacy management dimension on tags, we introduce here the concept of structured vocabularies. Thus, to each tag can correspond a structured vocabulary which positions the given tag in a hierarchy of related tags to specify a level of detail of the considered tag. A structured vocabulary defines an ordering in the tags

according to the semantic level of detail. We define than a structured vocabulary as a set of ordered tags:

$$\Psi(t_i) = \{t_i \in T, t_j \in T, t_i \neq t_j / t_i \prec t_j \vee t_j \prec t_j\} \tag{4.1}$$

It should be noted at this stage that each user may have her own structured vocabularies since users may associate different semantics to different tags and the granularity is not always defined with the same way. Also, communities of users may have a set of structured vocabularies after some agreement between members of the community (e.g. this is the case of Linked Data concept hierarchies, a result of a community effort in Wikipedia).

We define in this context the Social Privacy Pyramid (SPP), as a structure that attaches to each semantic concept (i.e., Tag) a set of other concepts with higher or lower granularity according to a social relations pyramid level.

Following this definition, a Dynamic Social Privacy Pyramid (DSPP) is a SPP constrained by an additional constraint, e.g., a user, used to translate the definition and the personalization of concepts to each user.

The main input for the approach we are proposing is the presence of a structured vocabulary associated to a tag. The particularity of this structured vocabulary is that it is intended to store the granularity of an information. That is, this structured vocabulary can be seen as a repository which stores a given information together with some of its different degrees of granularity (i.e. details).

We introduce in the next step a function ω defined as follows:

$$\omega : T \rightarrow \Psi(t_i),$$

$\mathrm{t}_j = \omega(t_i, \varepsilon) \in \Psi(t_i)(4.2)$

In other words, ω gives for an information or a tag t_i another form of it with more of less detail, say t_j, according to a relaxation parameter ε. We note by $\Psi(t_i)$ the set of all possible granular variations of a tag t_i.

The degree of granularity, controlled by the relaxation parameter ε, can be specified by the user for instance according to a set of predefined criteria.

An example of the use of the ω function is the following.

Consider the information "Birthday party" associated to a resource, in this case a photo which is uploaded to a social platform, say Facebook. An information with more degree of granularity is "Bob's Birthday Party" and one with less is "Social Event". Thus, as for the degree of granularity, Social Event precedes Birthday Party and Bob's Birthday Party succeeds Birthday Party.

We consider in the following that the owner of the photo specifies three social categories as criteria for degree of granularity: Friends, Family and Stranger. According to these preferences, ω show the degrees of granularity for each viewer category:

- ω (Birthday party, Friends) = Birthday Party
- ω (Birthday party, Family) = Private Social Event
- ω (Birthday party, Friends of Friends) = Social Event
- ω (Birthday party, Strangers) = Event

The input parameter for the precision level retrieved by the ω function can take a different form and this is based on the model of the social network. A possible different input parameter can take the form of quantified numerical values that annotate the social proximity value between nodes in the social network. In this case, Friends have a high value, while Strangers little. The example takes the following form:

- ω (Birthday party, $[50, 100]$) = Birthday Party
- ω (Birthday party, $[20, 50]$) = Private Social Event
- ω (Birthday party, $[10, 20]$) = Social Event
- ω (Birthday party, $[0, 10]$) = Event

In this case, it can be seen that there is an ordered relation between the tags, as for the degree of granularity: Birthday Party $\prec$ Private Social Event $\prec$ Social Event $\prec$ Event. From a social networking perspective, we consider that the user has a social network that is semantically rich, meaning that the relation between the user and another node in the network is annotated with the category (i.e. friend, family, professional, stranger) [25].

Thus, the role of these structured vocabularies is to provide a hierarchy of concepts for each tag category. This hierarchy will allow to select the level of detail to show for a specific tag according to a specific category of viewers. These structured vocabularies can be collaboratively built by a community or by the user. Existing vocabularies (ontologies, taxonomies) can also be reused if they correspond to the user's requirements. These structured vocabularies intend to propose a hierarchical view of a specific information, each level corresponding to a degree of granularity. More concretely, each level in the hierarchy corresponds to a level of detail.

In Figure 4.29 we show how this hierarchy allows to show different degrees of granularity of the same initial information, as a first proof of concept.

[25] Different approaches exist to have such a semantic social network but it is out of the main scope of this work to detail these techniques

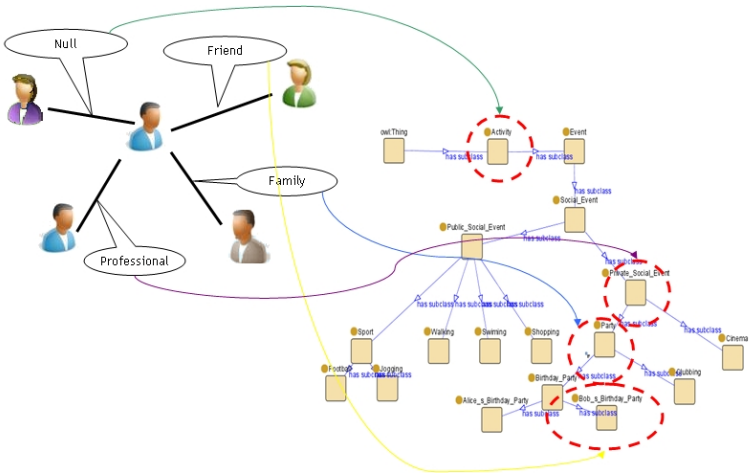


Figure 4.29: Proof of concept: different degrees of granularity of the event based on relationships

4.5.3 Granular Privacy Management Process

We discuss in this section what the proposed approach brings to the user by describing some principles that we follow to satisfy the constraints announced in the beginning of this work , i.e. reinforce the privacy on tags while keeping the tagging process as natural and as simple as possible.

As a first principle, we consider always that each tag inserted by the user is the most detailed tag independently from its position in a particular structured vocabulary. This is coherent with our consideration of the privacy problem where we aim at providing different levels of details for the same information w.r.t. different categories but this supposes that there doesn't exist more details than the given information from the user. As an example, consider the case of a photo in a social network. if Bob decides to annotate it with *Social Event*, this constitutes the origin of the information and we can not at any time provide for Bob's friends or family a more detailed one like *Birthday party*.

Even with the restrictive aspect that this first principle may impose to the use of such a method, this ensures a high privacy on the data. To relax this constraint, the structured vocabulary is used to provide suggestions for the user while she introduces the tags. This is helpful for the user to express as much details as possible for his close relative if she wishes.

The second principle that governs the approach we are proposing is the absence of a relation or logic between the number of categories and the levels of detail associated to tags in the structured vocabularies. In fact, the number of categories may be lower or higher than the number of different levels associated to a tag in a structured vocabulary.

The third principle is that different categories may be associated to the same level of detail. This reinforces the second principle and includes the case where the number of categories is too higher than the number of levels in a structured vocabulary.

The last principle of the approach is that the rights, i.e., degree of granularity, that a category has may be different from a resource to another. This means that there is no static order in the definition of the categories. More concretely, when a category c_1 with a high, say highest, degree of granularity on a tag t_1 associated to a resource r_1 may have the lowest degree on a tag t_2 associated to a resource r_2.

4.5.4 Implementation Proposal of the Granular Approach

4.5.4.1 System Architecture

From the architectural point of view, the main components of this solution are the following (Figure 4.30): (i) social network manager which is responsible of managing the access to the different social networks and gathering the social relations between the users. (ii) The category manager is responsible of keeping track of the different social categories that a user may have or has manually declared. (iii) The access rights manager is at the core of the proposal and is responsible of associating a privacy level to each category of social relations. The rules embedded into this component can be applied either on a category of users or a specific user. This explains the location of this component in the architecture. Finally, the (iv) user access rights manager configurator is a user interface which is used by the end user to introduce new rules, new categories, etc.

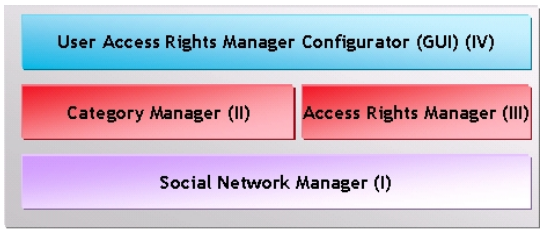


Figure 4.30: General architecture of the proposal

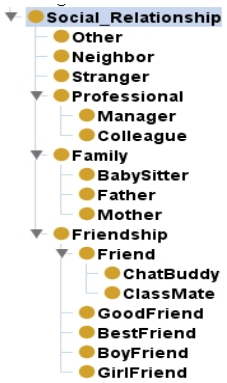


Figure 4.31: An example of the ontology which can control the mechanism

Finally, a mechanism that helps users to attach a specific detail level to pieces of information/resources needs to be used. This task can be performed manually, semi-automatically, or automatically. A manual way will involve the user too much and will

impose for the user to define everything herself for any piece of information. This is a very time and effort consuming task and not intuitive for the user. The semi-automatic way may exploit existing resources like ontologies and taxonomies (see Figure 4.31 for a very simple example of an ontology) which define generally different levels of the information like the category, sub-category, the concept itself, and eventually some reasoning rules to exploit the defined concepts, etc. but can involve the user in the final decision. An automatic process can be understood as a straight forward manner where information pieces are replaced with information of higher (or lower) level of detail. The automatic strategy may improve the efficiency but decreases the accuracy. The best strategy could be, from our point of view, the semi-automatic strategy where the user is involved in critical and ambiguous situations where the system is unable to decide by itself.

4.5.4.2 Example: Privacy in Social Network Conversation Spaces

We define in this scenario a conversation space in a social network, where people discuss in a dedicated space about a specific topic. Such a dedicated space is composed generally from people from different social spheres (friends, family, coworkers and strangers) and therefore it is important to consider the question of what information to show to a given person at a given time and in a given conversation. In other words, it is a subset of said social network composed of interactions[26] about a specific topic take place.

In the following, each member of such a discussion forum will be represented by their user profile, comprising a plurality of user profile attributes i.e. a set of items that represent their identity and interests.

The idea is to provide a user profile management that is able to help one user to blur attributes of his user profile depending on several parameters such as the purpose or subject of the discussion he wants to be connected to, his activity in the social network or his relationship with other people in this forum.

Figure 4.32 shows an example of how "John Glen"'s full profile information could be blured to members of a discussion about "Football".

As shown in Figure 4.32, the public profile manager uses an algorithm to show either an accurate or a more or less blured value of one profile attribute.

As will be described hereinafter, the method for managing a user profile within a social network will help a user to blur attributes of his user profile depending on several parameters such as the purpose or the subject of a subset of said social network he wants to be connected to, his activity in the subset, his relationship with other people in the subset etc.

Blurred profile attributes will therefore be provided to the members of said subset

[26] By interactions, one means any sort of activity conveying meaningful information between two or more people.

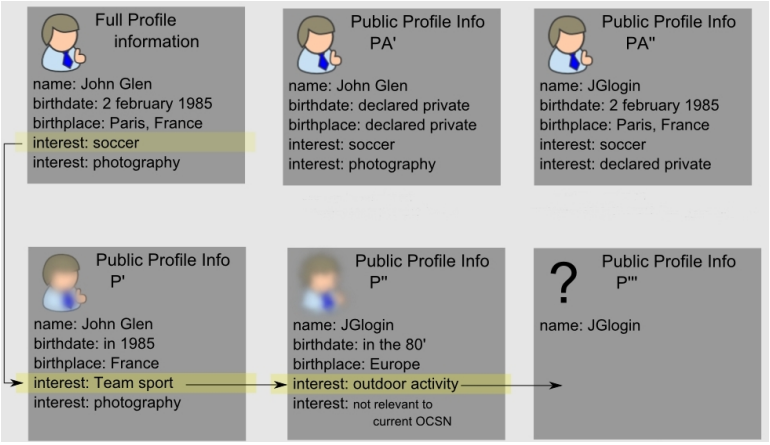


Figure 4.32: Bluring of user profile items

who have access to the public attributes or to any other third party having access to the public attributes, a third party being a physical person such as a member of the forum, an application, a service etc.

Main Steps of the Blurring Method

The method comprises:

- extracting a specific topic from interactions inside the subset of said social network
- for the topic, finding the most relevant matching semantic entity in a linked data graph (previous semantic matching algorithm)
- building a first directed graph of semantic entities for the most relevant matching semantic entity
- building a second directed graph of semantic entities for each said user profile attribute
- finding a first semantic entity inside the second directed graph of entity matching a semantic entity inside the first directed graph of semantic entities
- for each user profile attribute, displaying a second semantic entity from the second directed graph of entities according to the first semantic entity and at least one blurring criterion

The building of the first and second directed graph comprises:

- creating a starting node, said starting node being a semantic entity from the linked data graph and corresponding to said most relevant semantic entity
- creating other nodes representing semantic entities from the linked data graph, each pointed node being a semantic generalization entity of a source node, a pointed node and a source node being linked by an edge E, each edge E being directed and being of a unique type

For example, a user of a social network comprises the user profile with the following full profile information (attributes Att):

- Name(Att1) = "John Glen"
- Birthdate(Att2) = "2, February 1985"
- Birthplace(Att3) = "Paris, France"
- Interest1(Att4) = "Football"
- Interest2(Att5) = "photography"

Building the Directed Graphs

We consider in the following that the user is a member of a conversation space about "Sport". The question is what level of granularity to show about profile concept "Football" in this specific conversation space? The the following example considers the topic of the conversation space as the bluring criterion. As described afterwards, other bluring criteria can be considered also.

The propriety permits to consider the links that connect a semantic entity (here the starting node) to a semantic entity that expresses a more general information or the same information. A propriety Pr is associated to each semantic entity.

The following $SPARQL$ queries can be used. It allows retrieving all generalization semantic entities connected to a given specific semantic entity , say $C1$ "Sport".

The following query identifies the category concept associated to the query term:
$SELECT\ ?hasValue$
$WHERE$
$< http://dbpedia.org/resource/Sport >$
$< http://purl.org/dc/terms/subject > ?hasValue$

The second query retrives generalization concepts:
$SELECT\ ?prop\ ?value$
$WHERE$
$< http://dbpedia.org/resource/Category : Sports > skos : broader\ ?value$

From this query, one retrieves the semantic entities that have the starting node, here "sport" as a source node (S_Nod).

From the set of retrieved semantic entities (which are the neighbored semantic generalization entities to the starting node), those that are connected with the following propriety Pr which is a generalization link in the example given are considered.

The three following branches which represent the possible granular blurring schema for the most relevant semantic entity $C1$ can be retrieved with this query on keyword "sport" and corresponding category concept: "Category:Sports" (only a subset of possible results is shown):

- Sport - Category:Sports
- Category:Sports«Category:Exercise«Category:Health_effectors«Category:Health
- Category:Sports«Category:Games
- Category:Sports«Category:Recreation«Category:Hobbies«Category:Personal_life

Hence, for the 1st branch, the other nodes will be, Category:exercise, Category:health_effectors, Category:health and the edges will be:

- E1: Category:Sports - Category:Exercise
- E2: Category:Exercise - Category:health_effectors
- E3: Category:health_effectors - Category:health

For the 2nd branch, the other nodes will be ball_games and games and the edges will be:

- E1: Category:Sports - Category:Games (S_Nod-Nod1)
- E2: Category:Games - Category:Recreation (Nod1-Nod2)

For the 3rd branch, the other nodes will be Category:recreation, Category:hobbies, and Category:personal_life and the edges will be:

- E1: Category:Sports - Category:recreation (S_Nod-Nod1)
- E2: Category:recreation - Category:hobbies (Nod1-Nod2)
- E3: Category:hobbies - Category:personal_life (Nod2-Nod3)

Consider in the following the user profile. The same operation is performed on a given user profile attribute, say an interest in "Football".

In this above none-limitative example, 2 branches are retrieved. For the 1st branch, the other nodes will be Category:Ball_games, games, hobbies, personal_life, self, humans and the edges will be:

- E1: Football - Category:Ball_games (S_Nod-Nod1)
- E2: Category:Ball_games - Category:Sports_by_type (Nod1-Nod2)
- E3: Category:Sports_by_type - Category:Sports (Nod2-Nod3)
- E4: Category:Sports - Category:Recreation (Nod3-Nod4) etc.

For the 2nd branch, an example of the the other nodes will be Tteam_Sports, Category:Outdoor_Activity, Category:Collaboration, Category:Human_Behavior, Category:Humans and the edges will be:

- E1: Football - Category:Football (S_Nod-Nod1)
- E2: Category:Football - Category:Team_sports (Nod1-Nod2)
- E3: Category:Team_Sports - Category:Collaboration (Nod2-Nod3)
- E4: Category:Collaboration - Category:Human_Behaviour (Nod3-Nod4)
- E5: Category:Human_Behaviour - Category:Humans (Nod4-Nod5)

Computation of Bluring Level for the Profile Attribute

A first semantic entity is found inside the second directed graph of entity that matches a semantic entity inside the first directed graph. Hence, the first semantic entity is the less abstract concept which distance with the starting node of the second direct graph is the lowest.

In the example given, the semantic entity "Ball_Games" is found matching the most relevant semantic entity of the first directed graph. Therefore, instead of "Football", the entity "Ball_Games" will be shown, as the closest entity with minimal distance to the attribute node and topic node and present in both directed graphs.

An example SPARQL query performing this computation is shown below (searching for the most common concept between concepts "Football" and "Sport":
$PREFIX\ db : < http : //dbpedia.org/resource/ >$
$PREFIX\ rdf : < http : //www.w3.org/1999/02/22 - rdf - syntax - ns\# >$
$PREFIX\ skos : < http : //www.w3.org/2004/02/skos/core\# >$
$SELECT * WHERE$
$db : Sport\ ?pf1\ ?middle.$
$db : Football\ ?ps1\ ?os1.$
$?os1\ ?ps2\ ?os2.$
$?os2\ ?ps3\ ?middle.$
$FILTER\ (($
$(?pf1 = < http : //purl.org/dc/terms/subject >)$

$||$
$(?pf1 = skos : broader)$
)

(
$(?ps1 = < http : //purl.org/dc/terms/subject >)$
$||$
$(?ps1 = skos : broader\ )$
)

(
$(?ps2 = < http : //purl.org/dc/terms/subject >)$
$||$
$(?ps2 = skos : broader\ )$
)

(
$(?ps3 = < http : //purl.org/dc/terms/subject >)$
$||$
$(?ps3 = skos : broader\ )$
)
).

Example of result of this query, showing the minimal path between the two aformentioned concepts (Fig.).

http://purl.org/dc/terms/subject -
http://dbpedia.org/resource/Category:Sports -
http://purl.org/dc/terms/subject -
http://dbpedia.org/resource/Category:Ball_games -
http://www.w3.org/2004/02/skos/core#broader -
http://dbpedia.org/resource/Category:Sports_by_type -
http://www.w3.org/2004/02/skos/core#broader

In conclusion, concept "Ball_Games" will be shown instead of Football.

4.5.5 Management of Bluring Criteria

Different non-limitative examples are given below for these criteria.

- If the subject of the subset of the social network (e.g. conversation space) is very similar to the object of the attribute *Att*, the level of blur for the value of the attribute *Att*4 will be set to a low level, which means that a more accurate value

of the attribute *Att* can be displayed than for other attributes *Att*. Hence, in the example given, as the subject of the subset is "sports" and the attribute *Att*4 in the full profile information is "Football", the second semantic entity AC disclosed to the members of said subset will be first semantic entity LAC i.e. "ball_games" with a distance $D1 = 1$. In order to check the similarity, one selects the most relevant semantic entity in the second directed graph Dgc of the attribute *Att* which corresponds to the subject of the subset. The distance D between the most relevant entity selected and the starting node $S - Nod$ is retrieved and the level of blur is set to that distance D.

- If the owner X of the user profile and user Y already discussed a lot in the past (they had many interactions), user X will likely want to disclose to user Y more accurate values from his full profile information. A low level will be assigned to the level of blur LB and the second semantic entity AC which has a low semantic distance from the first semantic entity LAC will be disclosed, for example "ball_games" or "team_sport".

  Otherwise, the second semantic entity LAC which has a high semantic distance D from the second semantic entity AC will be disclosed, for example "outdoor_activity" or "collaboration".

  In a non-limitative example, proximity exploitation may be performed using predefined thresholds. For example, some predefined thresholds may be defined for the level of blur LB according to the number of discussions engaged between the owner X and a user Y. Hence, for example, if the discussions are over 100, LB = 0, if the discussions are over 50, LB=1 etc.

- If the owner of the user profile is an important contributor to the subset (he has/had many discussions with many people for instance) or has a lot of influence of this community, members of the subset could access more accurate value of the attributes related to the subject of the subset than for other attributes.

  Some predefined thresholds may be defined for the level of blur LB according to the number of contributions (number of messages post for example) of the owner.

  Some predefined thresholds may be defined for the level of blur LB according to the influence of the owner (number of messages originated from the owner which have been broadcasted by the other members, number of messages of the owner scored).

- The predetermined level blurring level LB is set by the owner of the user profile when creating his profile for example

- Blurring strategy may be configured in advance in order to ease the selection by the owner or to automatically choose the level of blur for each attribute in the full

profile information. Hence, for example, the system will provide several strategy of privacy management. In non-limitative example, four strategies are proposed:

- Strategy 1 where all the attributes *Att* declared public to everyone will be disclosed with their accurate values. Therefore, all *Att*1 to *Att*4 will be disclosed.
- Strategy 2 where only the attributes *Att* relevant to the subset are disclosed with the accurate values. Therefore, only *Att*4 will be disclosed with the value (semantic entity) "ball_games".
- Strategy 3 where all the attributes *Att* declared public to everyone will be disclosed with lightly blurred values.
- Strategy 4 where all the attributes *Att* declared public to everyone will be disclosed with heavily blurred values.
- Strategy 5 where all the attributes *Att* declared public to everyone will be disclosed with heavily blurred values except for the users who had many interactions in the past on this subset with the owner X of the user profile.

4.6 Summary of the Contributions

In this chapter we introduced our semantic framework for social search from a theoretical point of view by focusing on the algorithms that perform the analysis of microposts. In a second part, we address the issue of user profiling based on Linked Data concepts and we propose an approach for a more convenient way of managing the privacy in such a system.

The analysis of microposts is achieved by a toolkit of content processing algorithms, built on top of a state-of-the-art keyword/named entity extractor. The first algorithm connects the extracted keywords and named entities to concepts from a semantic knowledge base. Currently the algorithm is specifically adapted to work with DBPedia. The second algorithm deals with the issue of semantic expansion of concepts in order to further approximate user interests. Once the conceptual identification of user interests is achieved, the second step consists of scoring the user prorile concepts in order to evaluate the level of expertise. In this case, we take into account only information about the semantics of the shared messages and not the structure of the network. Our objective is not to further increase the power-law, by recommending people who have lots of connections, but to identify people who share interesting content in a given domain.

The following chapter will present a proof-of-concept application that is built on top of this framework, performing social search on top of Twitter, a popular social network.

Chapter 5

Implementation and Evaluation of the Framework

Contents

In the previous section, we introduced the main theoretical contributions of the framework, mainly in the area of transforming keywords in microposts into semantic concepts. This included the theoretical framework, the overall system architecture and the specific algorithms that perform the individual tasks for the user profile construction, content extraction and social search. In this section we present an application, available online, called the "Tagging Beak", which acts as a proof-of-concept for the theoretical framework. The Tagging Beak is an online available `www.tbeak.com` social search engine on top of Twitter. It implements the social search and socio-semantic awareness in the form of a Q&A system.

Practically, users can ask questions in natural language. The system applies the developed toolkit on each question and constructs a list of recommended users who are the most relevant regarding the topic(s) of the question. In other words, a dynamic

community is constructed that is composed of users with high expertise on the specific subject areas. Therefore, the Tagging Beak facilitates the creation of seamless, trusted interaction spaces between members of a community.

5.1 Implementation of the Social Search Framework

The framework is constructed on top of a real social platform, Twitter, and integrates the algorithms mentioned in the previous chapter. From a technical point of view, the system was implemented in the open source framework Drupal 6 [1] and integrates two data stores: (i) a relational database populated by the crawler engine and (ii) a semantic database which uses the data model of MOAT (Meaning of a Tag) for storing the processed social information.

In the following section, we focus on the implementation of the crawler engine, which is responsible of extracting the social interactions from the social platform. It is to be noted that Twitter can be replaced in the framework by another social platform or they can be combined. For this, the corresponding API has to be implemented (e.g. OpenSocial or Facebook API). Their description can be found in the Appendix.

5.1.1 The Crawler Engine: Crawling social interactions using the Twitter API

The Tagging Beak implements the content capture layer using a crawler on Twitter. The content capture layer's objective is to connect to a social platform and capture the social interactions shared by a given user and members connected to said user. In the current implementation of our framework, Twitter is used as a source social platform[2]. The current implementation of the Facebook API and that of OpenSocial API makes it difficult to work with social data from Facebook or Netlog from the point of view of our research mainly because of scalability issues - e.g. in Facebook in order to extract content, the explicit authorization of the user is necessary [4].

The layer allows to connect other social platforms and no modification is required in the database structure. Only the corresponding crawler specific to the social platform must be implemented. As a source of social data, Twitter is particularly interesting because of its widespread media coverage. Social interactions on the site (i.e. Tweets) are extremely short, can be threaded, and for the most part are publicly visible.

[1] Drupal Open Source Content Management System - `http://drupal.org/` - visited August 2011

[2] We performed an exploratory analysis regarding the potential integration of Facebook and systems based on OpenSocial[3]

[4] The results of this analysis can be found in the Appendix of this document, chapter "The case of Facebook". OpenSocial based platforms are more promising, as shows the corresponding analysis. However, for the need of our research, Twitter is sufficient as it provides scalable, rich and real-time social data

```
    "user": {
      "profile_sidebar_border_color": "87bc44",
      "name": "Twitter API",
      "profile_sidebar_fill_color": "e0ff92",
      "profile_background_tile": false,
      "profile_image_url":
"http://a1.twimg.com/profile_images/689684365/api_normal.png",
      "created_at": "Wed May 23 06:01:13 +0000 2007",
      "location": "San Francisco, CA",
      "profile_link_color": "0000ff",
      "follow_request_sent": false,
      "favourites_count": 5,
      "url": "http://apiwiki.twitter.com",
      "contributors_enabled": true,
      "utc_offset": -28800,
      "id": 6253282,
      "listed_count": 4591,
      "profile_use_background_image": true,
      "followers_count": 196156,
      "lang": "en",
      "protected": false,
      "profile_text_color": "000000",
      "time_zone": "Pacific Time (US & Canada)",
      "notifications": false,
      "verified": true,
      "geo_enabled": true,
      "profile_background_color": "c1dfee",
      "description": "The Real Twitter API. I tweet about API
changes, service issues and happily answer questions about
Twitter and our API. Don't get an answer? It's on my website.",
      "statuses_count": 1955,
      "profile_background_image_url":
"http://a3.twimg.com/profile_background_images/59931895/twittera
pi-background-new.png",
      "friends_count": 20,
      "show_all_inline_media": false,
      "screen_name": "twitterapi",
      "following": true

    }
```


Figure 5.1: Example Twitter user data in JSON format

5.1.1.1 Available data and permissions in Twitter

Figure 5.1 shows an example of the data available for a Twitter user in JSON format (XML, RSS and Atom are also available).

In addition to this information, the Twitter API provides the friends of a Twitter user (i.e. the other users that they are following) and the followers of a Twitter user. All of this information is normally publicly available to anyone that knows a screen name or user ID. However, Twitter users can opt to have a "protected profile", meaning that only their followers can view any of their information. Unlike a public account, followers to a protected profile have to be approved by the Twitter user. Any call to the Twitter API is made in the name of an authenticated user, and will contain information from protected profiles if that user has permission to see it. Figure 5.2 shows an example of what information can be obtained about a tweet. Again, all tweets are considered public unless the source user has a protected profile. Much of the information is optional,

```
{
    "coordinates": null,
    "created_at": "Thu Jul 29 16:04:42 +0000 2010",
    "truncated": false,
    "favorited": false,
    "text": "There are 26 days left to migrate from basic auth
to OAuth. http://t.co/fhQMvSj Need help? Mailing list:
http://t.co/CmUtcG5 ^TS",
    "contributors": [819797],
    "id": 19836772921,
    "geo": null,
    "in_reply_to_user_id": null,
    "user": {... SEE Figure 2 ...},
    "source": "web",
    "place": {
      "name": "SoMa",
      "country": "The United States of America",
      "country_code": "US",
      "attributes": { },
      "url":
"http://api.twitter.com/1/geo/id/2b6ff8c22edd9576.json",
      "id": "2b6ff8c22edd9576",
      "bounding_box": {
        "coordinates": [
          [[-122.42284884, 37.76893497],
           [-122.3964, 37.76893497],
           [-122.3964,37.78752897],
           [-122.42284884, 37.78752897]
          ]
        ],
        "type": "Polygon"
      },
      "full_name": "SoMa, San Francisco",
      "place_type": "neighborhood"
    },
    "in_reply_to_screen_name": null,
    "in_reply_to_status_id": null
}
```


Figure 5.2: Example tweet in JSON format

especially information about geolocation.

The Twitter API is fairly comprehensive, and permits more active changes to the Twitter system, such as posting a Tweet, modifying the user's profile, following/unfollowing other users, marking other users as spam, etc.

5.1.1.2 Obtaining Credentials in Twitter

This section focuses on using one (or more) Twitter users to crawl publicly available tweets. This is sufficient for the needs of a generic social data crawler. For a system that requires more personalized, private and semi-private information from Twitter (tweets from protected accounts, direct messages, and taking action on behalf of a user), further implementation is required. For example, applications built on top of this framework can be web-based interface that acts as a front end to the user's personal social data,

and integrates with the social data crawler described here.

The OAuth protocol [5] is used to connect through the Twitter API, and requires credentials for two actors. The application obtains a consumer key and secret through the Twitter system [6]. To use the Twitter crawler, a new application should be registered with Twitter and the consumer key and secret specified in a "TwitterCrawler.xml" configuration file.

All API requests are made in the name of a specific Twitter user, in the context of an application, and that user needs to obtain an access token and secret. For the Twitter crawler, a new user can be created to act on its behalf. It is a good idea to request whitelisting [7] from the system to increase the API limits (to approximately 20,000 requests an hour). Note that if an existing user is used, the private and semi-private information from that user will be available to the Twitter crawler.

The Twitter crawler includes a mechanism for out-of-band (or PIN) authentication to help generate the requesting key/secret. Running the Twitter crawler without an access key/secret pair will generate an command line prompt with instructions on how to generate and configure the access key/secret pair.

5.1.1.3 Generic Implementation of a Twitter Crawler

The Twitter crawler relies on the framework's content capture component to determine which users (and their friends) to crawl. The user must have declared a Twitter account that has been activated for crawling. The Twitter crawler searches for three types of information:

- Interactions: the tweets that the user has published
- Profiles: the personal information that the user has saved on Twitter, and
- Relations: the friends/followers associated with the user.

Each type of information can be enabled/disabled independently of each other according to the configuration of the crawler 5.1. For example, the parameter *twitterProcessInteractionRate* determines how frequently the crawler makes API calls to find new interactions (tweets) by specifying the number of seconds to wait between calls. Each API call returns up to 20 new tweets for one specific user, so the default value of 1.0 means that the system can process at most 3600 users and 72000 tweets in one hour. A negative value disables crawling tweets.

After crawling, the interaction information is saved in a normalized interaction model managed by the content capture component.

[5] The OAuth 2.0 Protocol, draft-ietf-oauth-v2-10 (http://tools.ietf.org/html/draft-ietf-oauth-v2-10, visited the 2010/07/28

[6] http://twitter.com/apps

[7] http://twitter.com/help/requestwhitelisting

Name	Description	Default value
oauthConsumerKey	oauthConsumerKey	n/a
oauthConsumerSecret	The OAuth consumer secret for our Twitter indexing application.	n/a
oauthAccessToken	The OAuth access token for the user making Twitter API calls.	n/a
oauthAccessSecret	The OAuth access token secret for the user making Twitter API calls.	n/a
twitterProcessInteractionRate	The period at which the tweets are crawled (in seconds).	1.0
twitterProcessProfileRate	The period at which user profiles are crawled (in seconds).	1.0
twitterProcessProfileStaleness	The minimum period to wait before refreshing user profile information (in seconds).	1200.0
twitterProcessRelationRate	The period at which user relationship information API calls are made (in seconds).	1.0
autoCreateUnknownTwitterUsers	Add previously unknown Twitter users to the system when they are encountered	true
crawlAutocreatedUsers	Start crawling autocreated Twitter users.	false

Table 5.1: TwitterCrawler configuration

Likewise, the parameters *twitterProcessProfileRate* and *twitterProcessRelationRate* determines how frequently the crawler makes API calls to find profile information and/or relationship information to be stored in the content capture component. The *twitterProcessProfileStaleness* parameter is used to ensure that API calls aren't used to re-fetch the same user's profile information, especially in demonstration situations. This information is also saved in the content capture component. The Twitter crawler cycles through the list of users with activated Twitter accounts, so that the number of users affects how frequently their information is indexed. The two parameters *autoCreateUnknownTwitterUsers* and *crawlAutocreatedUsers* are used to determine how interactions with non-users are treated. This is important to the analysis component.

For example, Alice (a user with an activated Twitter account) sends the tweet "@Bob How was your #vacation?" Note the conventions for a destination (the Twitter user "Bob") and a hash tag (the keyword "#vacation"). If Bob is known to the system (i.e. he is also a user in the system with a Twitter account), the analysis server applies the extracted information to the relationship between the two users. If Bob is not in the system, the tweet can only affect Alice, perhaps incrementing the importance of the hash tag "vacation" to her.

The *autoCreateUnknownTwitterUsers* parameter, if true (as by default), causes the system to ensure that a user exists or is created for each Twitter user encountered, associated with the discovered Twitter account, permitting relationship information to be constructed for every user. However, the auto-created Twitter accounts are not crawled unless the crawlAutocreatedUsers is also true (false by default). However, because of the interconnected nature of Twitter, this can result in an explosion in the number of users being crawled.

5.1.1.4 Limitations

Each Twitter API call is rate-limited by requesting user and/or IP address. A whitelisted user has 20,000 API calls/hour. The default configuration will never make more than 10,800 per hour (and typically less, since that figure doesn't include the processing time for each call). It should be safe to reduce the three *twitterProcessXxxRate* parameters to 0.5 seconds and remain within the limits (or any combination where their sum is >= 1.5).

5.1.2 Crawling Engine Statistics

This section presents the current dataset (as for 30 August 2011) crawled in the prototype application. This data has been gathered in a period of 2 Months.

- Total Tweets: 4502690
- Total Accounts: 18076

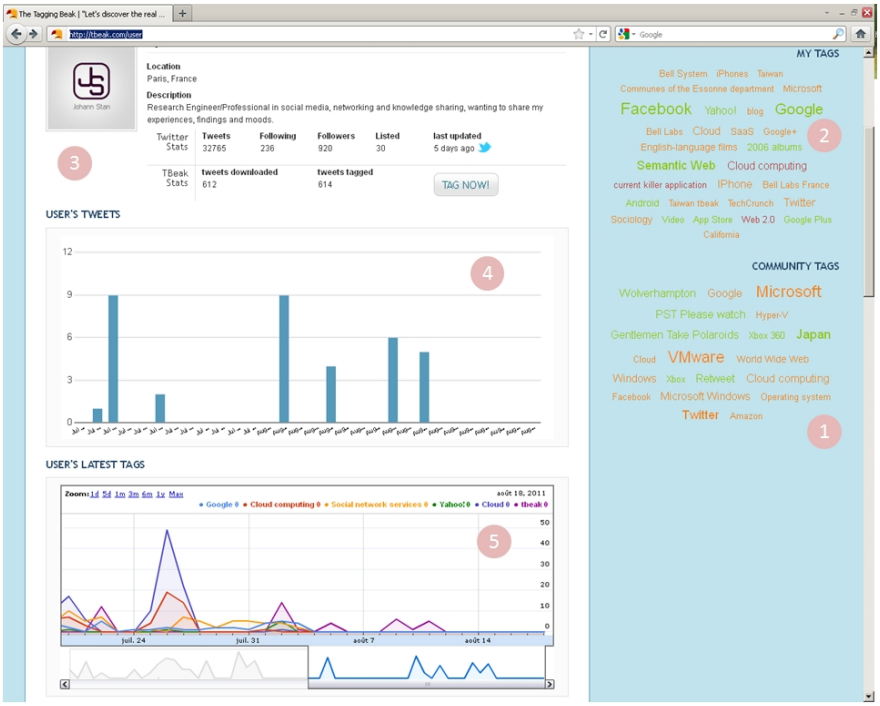


Figure 5.3: Upper-View of the Tagging Beak: interface of the user profile

- Total Hashtags: 1774335
- Total URL: 2731253
- Total User mentions: 3337529
- Total ReTweets: 411757

5.1.3 User Interface

The Tagging Beak allows users to connect with their existing Twitter account. Once connected, the profile construction process can be executed, which analyses the shared messages by applying the social semantic matching (SoSeM) and semantic expansion (SemEx) algorithms. In a second step, the profile scoring process will associate the weights to each profile item and perform the propagation of scores to expanded concepts. The result of this phase is a user expertise interaction profile in form of a colored tag cloud, shown in Figure 5.3.

As seen in Figure 5.3, the user profile interface is composed of several elements:

1. *Community Tags:* most frequent concepts shared by the community of the user.

2. *My Tags:* the expertise profile of the user (colors represent the average sentiment associated to each concept)

3. General information about the user

4. The interactivity of the user: number of shared messages / day

5. Relative interactivity of the user in the case of the most frequent concepts

6. Detailed information about the results of the analysis on each micropost

5.1.4 Social Search Interface

The main feature of the application is the social search. This is implemented in the following form. Users can ask questions in natural language. The system will show a list of people from the user's community who are relevant in expertise to the topics of the question. Figure 5.4 shows the general interface of the social search component, composed of:

1. A frame where users can express their information need in natural language (e.g. a question, a message, keywords etc.) - (1)

2. The semantic processing of the said information need using the algorithms $SoSeM$ and $SemEx$ - (2). The Figure shows the extracted and expanded concepts.

3. The construction of a list of people from the social network of the user relevant to said information need (i.e. similarity of profile concepts) and their ranking according to expertise and interactivity - (3). Profile information can be visualized about each person. The granular privacy approach is used to compute what profile attributes to show and with what level of granularity, according to the topic of the question.

4. Several interaction modalities offered to the user to engage into a conversation with said people. This can be an explicit interaction, implemented as the possibility to directly ask the said user or a more implicit interaction, consisting of following the recommended user - (4)

5.

In addition, Figure 5.5 shows an upper-view of the analysis of a micropost, including the application of knowledge extraction algorithms as well as showing the corresponding sentiment polarity score. We included in the knowledge extraction process also tags that are extracted from the content of the hyperlink. Also, the expansion of these concepts is shown in the Figure.

Figure 5.4: Upper-View of the Tagging Beak: social search interface

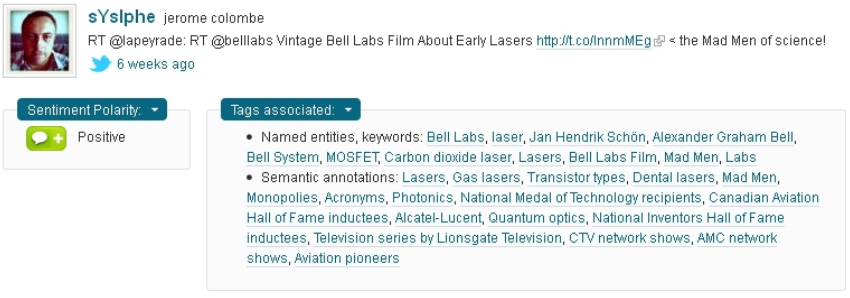


Figure 5.5: Upper-View of the analysis of a micropost

5.2 Evaluation of the Algorithms

5.2.1 General Evaluation Protocol: User Study

In this section, we specify the requirements of the evaluation, describe the experimentation plan and setup we followed, then discuss the results obtained.

The evaluation of our approach relies on the quality of two data processing steps: (i) the extraction of weighted concepts from microposts, and (ii) the construction of semantic tag clouds representing a contextual expertise profiles for each user, composed of concepts more recent than one week.

As recommended by [Terveen 2005], "Evaluations of social matching systems should focus on users and their goals", we decided not to rely on existing evaluation data sets such as the ones from TREC, nor to follow a scenario-based experiment. Instead, we leverage usual browsing behavior of the users, and their own social updates, with their consent during the period of the experiment.

The scores expected from volunteers are threefold: (i) the validity of explicitly shared profile concepts and the added-value of context in the disambiguation process, (ii) the validity of expanded profile concepts (from both a knowledge base and URLs present in the micropost), and (iii) the relevance of recommended users to a question in the case of (i) keyword-based and concept-based ranking function.

5.2.2 Experimentation of the Semantic Matching Algorithm

The experimentation protocol for the semantic matching algorithm is described as follows. The objective of this experimentation is to verify whether the auto-tagging process successfully matched the keyword to the relevant DBpedia concepts. More concretely, we compare the baseline approach, which uses only the named entity without context with our approach that includes additional contextual cues (i.e. the user's profile).

The protocol of experimentation is then integrated in the auto-tagging process, more precisely at the moment when the system tags the user's tweets. Each time, the system finds a keyword, and matches it to a concept, the informations about this process (i.e. user ID, the content of the tweet, the keyword found, the concept matched and an abstract of this concept) are inserted into the table "experimentation_sm" in the database. The objective is to query this table to retrieve all the couple keyword-concept tagged by the system to evaluate.

The informations to be saved:

- Name: screen name of user's twitter account
- Message: the message contains the keyword
- Keyword: the keyword found in the message

- Concept: the concept found for that keyword
- Abstract: the abstract of the found concept

We then query this base to create a form in which we present all of these information and an additional column "Score".

These forms were automatically generated and people could access it in the prototype in order to annotate the concepts for their messages. The score is evaluated from 1 to 5 depending on its relevancy to the keyword and the message:

- 5 if the concept corresponds exactly to the concept
- 3 or 4 if the concept is not exactly the keyword but have relation to the message or the user's context
- 1 or 2 if the concept is related to the user's contexts (e.g. stays in the same category or the same topic of interest)
- 0 if the concept has no relation neither to the keyword nor the message or the user's context

After having the evaluation, we compute the precision of the algorithm by the formula ($score_{max} = 5$):

$$Precision = \frac{\sum_{i=1}^{N} score_i}{score_{max}.N}$$

From the informations saved during the experimentation, we have the following database view for a randomly selected user profile (Fig. 5.6. In this experiment, 12 volunteers participated with different backgrounds (i.e. 6 students, 2 researchers and 4 Twitter users from the Semantic Web community).

Our experimentation returned the following scores in average for each profile (Figure 5.7):

- Number of total concepts found: 94 (all the repetitions are removed).
- Number of concepts rated exactly correct (5 points): 58/94 (i.e. 61,7%)
- Number of concepts that got 4 points: 8/94
- Number of concepts that got 3 points: 6/94
- Number of concepts that got 2 points: 5/94
- Number of concepts that got 1 points: 9/94

expe_name	expe_message	expe_keyword	expe_concept	expe_abstract	expe_score
socialhead	RT @workinthecloud: #SaaS Why SaaS HR makes Busine...	saas	http://dbpedia.org/resource/Saas-Almagell	Saas-Almagell is a municipality in the district of...	0
socialhead	RT @workinthecloud: #Cloud iCloud is Apple's chanc...	apple	http://dbpedia.org/resource/Apple_Inc.	Apple Inc. is an American multinational corporatio...	0
socialhead	RT @workinthecloud: #Cloud iCloud is Apple's chanc...	cloud	http://dbpedia.org/resource/Cloud_computing	Cloud computing is Internet-based computing, where...	0
socialhead	RT @workinthecloud: #Cloud Cloud Commerce Focuses ...	cloud	http://dbpedia.org/resource/Cloud_computing	Cloud computing is Internet-based computing, where...	0
socialhead	RT @workinthecloud: #Cloud Cloud Commerce Focuses ...	small	http://dbpedia.org/resource/Birmingham_City_F.C.	Birmingham City Football Club is a professional as...	0
socialhead	RT @workinthecloud: #Cloud IBM Brings High-perform...	ibm	http://dbpedia.org/resource/IBM	International Business Machines (IBM) is an Americ...	0
socialhead	RT @workinthecloud: #Cloud IBM Brings High-perform...	high-performance computing	http://dbpedia.org/resource/High_Performance_Compu...	The High Performance Computing and Communication A...	0
socialhead	RT @workinthecloud: #Cloud IBM Brings High-perform...	cloud	http://dbpedia.org/resource/Cloud_computing	Cloud computing is Internet-based computing, where...	0
socialhead	RT @workinthecloud: #SaaS 'Granite Horizon In The ...	pr newswire	http://dbpedia.org/resource/PR_Newswire	PR Newswire started out in 1954 as a vendor hired ...	0
socialhead	RT @workinthecloud: #SaaS 'Granite Horizon In The ...	pr newswire	http://dbpedia.org/resource/Xinhua_PR_Newswire	Xinhua PR Newswire is a venture created by Xinhua ...	0

Figure 5.6: Experimentation table. Database administrator view.

- Number of wrong concepts : 8/94

Precision value:

$$Precision = \frac{58.5+8.4+6.3+5.2+9.1+8.0}{5.94}\% = 75,3829\%$$

One interesting thing is 75% is not the maximum precision of the algorithm because at the moment of the experimentation the system had no existing context and with a hundred of concepts found, the context is not very rich either. Therefor during the similarity computing process, to not loose any potential concept, not only one concept but several concepts are returned and this mechanism at first gives bad scores to the precision because it adds also the non-relevant concept. However, after a period of time when the context is rich, we have only one correct concept added to the context and that improve rapidly the precision score.

An interesting observation during this evaluation was the possibility to cluster the vocabulary constructed from the contextual cues into sub-vocabularies specific to a given context of the user. Specifically, when sharing about "Apple" at work, this may refer to the company, but not when the user is in a different context, i.e. at home. Thus, giving more weight to part of the vocabulary that originates from the same context as the current message to be disambiguated could further improve the precision of the disambiguation.

A second observation with regards to the precision is the fact that recently shared terms in the vocabulary should have more weight in the disambiguation process. This

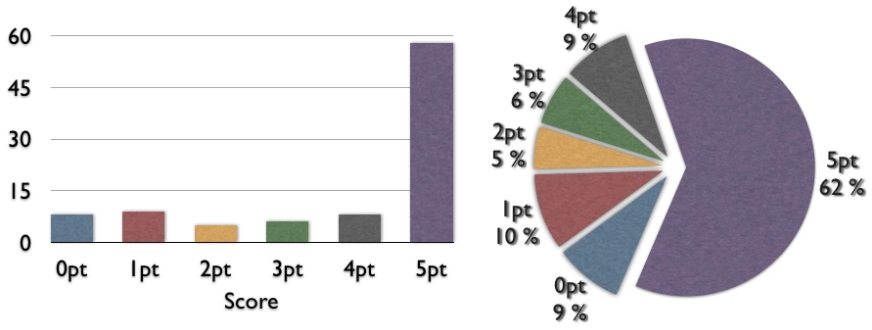


Figure 5.7: Experimentation statistics. Distribution of evaluation scores.

could play an important role in the situation where the user shared the same amount of messages about a given concept, say "Apple" and with two different meanings.

These improvements can be easily added to the algorithm and could work well in specific scenarios. However, our tests on the collection of microposts showed that such scenarios occur very rarely and the context of the message together with the vocabulary of previous messages as well as the community seemed sufficient to match to the right concepts with high precision. Therefore, these improvements are suggested as future work and are not included in the current version of the algorithm.

In order to make the experimentation more interesting, we compare the result of our algorithm in two cases: the context construction activated and not.

The following table present the keyword and concept that the Social Semantic Matching (SoSeM) algorithm found for a specific Twitter user (Figure 5.8

Nb	Keyword	Alchemy result	SoSeM result
1	Antwerpen	null	http://dbpedia.org/resource/Antwerpen-Centraal_railway_station
2	Antwerp	null	http://dbpedia.org/resource/Antwerp
3	Bruxelles	null	http://dbpedia.org/resource/Groupe_Bruxelles_Lambert
4	Belgium	http://dbpedia.org/resource/Belgium	http://dbpedia.org/resource/Belgium http://dbpedia.org/resource/Leopold_II_of_Belgium http://dbpedia.org/resource/Limburg_%28Belgium%29
5	weekend	null	http://dbpedia.org/resource/Weekend_Edition
6	Anvers	null	http://dbpedia.org/resource/Anvers_%28Paris_M%C3%A9tro%29
7	Paris	null	http://dbpedia.org/resource/Paris
8	Clint Eastwood	http://dbpedia.org/resource/Clint_Eastwood	http://dbpedia.org/resource/Clint_Eastwood
9	Gran Torino	null	http://dbpedia.org/resource/Gran_Torino_%28film%29
10	great movie	null	http://dbpedia.org/resource/The_Great_Movie_Ride
11	http	null	http://dbpedia.org/resource/HTTP_referrer http://dbpedia.org/resource/Apache_HTTP_Server
12	interesting	null	http://dbpedia.org/resource/Interesting_Times

Figure 5.8: Alchemy vs SoSeM. Social Semantic Matching on "BobMovie" twitter account.

We compare the concepts found in the two situations (the first case if called Alchemy in the database) for the last 20 tweets of a twitter account (i.e. "bobmovie" account, this is a test account created to tweet about movie interests in an ambiguos way (i.e. few contest available). The main objective is to disambiguate the keyword "Gran Torino" which is ambiguous between the name of a movie and the name of a car.) The result shows that when there is no context associated, the algorithm found only 2/12 concepts (not including Gran Torino) while in the second case algorithm found all 12/12 the concept and 6/12 are exactly correct (concept 1, 2, 6, 7, 8, 9), the others contain the correct concept but also contain several concepts at the same time (because of the context has not been stable yet).

In addition, our algorithm disambiguated successfully the keyword "Gran Torino" by matching it to the movie concept, which shows that the context of previous messages can help in this process [8].

The second comparison is realized on the twitter account, this account contains the tweets that are essentially about *cloud computing*, *telecommunication* and *news in information technologies*. The result between situations where context is used and not validate again the advantage of using context in the disambiguation process of microposts, almost all the keywords are matched to the correct concept (Figure 5.9.

5.2.3 Semantic Expansion Algorithm Evaluation

5.2.3.1 Experimentation protocol

The semantic expansion as we said before is the last step in the process of the User Profile construction. It means that after this step, we are able to build for each user a profile. This profile contains the concepts extracted from his/her own posts in the community and the other concepts expanded from those concepts. To validate the semantic expansion comes back actually to evaluate the user profile if it is well constructed. To do that we base on the same idea with the experimentation we used on the Social Semantic Matching algorithm.

We invite users to evaluate his/her own profile, meaning to give a score to each concept present in his/her profile. The evaluation score is then stored and the precision is computed with the same formula as the previous ones. We then also compare this precision to a threshold to be able to decide to validate or not the expansion. An example of this process in Figure 5.10 show the expansion of concept "Facebook":

The experimentation for the semantic expansion has been conducted with 12 voluntary participants (mentioned before) in Twitter who were asked to log in and then rate the concepts found in their profile.

Our first objective was to understand what kind of expansion is mostly appreciated by users (e.g. hierarchical, concepts from related categories or more directly connected

[8] Gran Torino Movie - `http://dbpedia.org/page/Gran_Torino_%28film%29` - visited June 2011

Nb	Keyword	Alchemy	SoSeM
1	amazon	http://dbpedia.org/resource/Amazon.com	http://dbpedia.org/resource/Amazon.com
2			http://dbpedia.org/resource/Amazon_River
3	apple	http://dbpedia.org/resource/Apple_Inc.	http://dbpedia.org/resource/Apple_Inc.
4	bell labs	http://dbpedia.org/resource/Bell_Labs	http://dbpedia.org/resource/Bell_Labs
5	business review	null	http://dbpedia.org/resource/Business_Review_Weekly
6			http://dbpedia.org/resource/Harvard_Business_Review
7	carleton university	http://dbpedia.org/resource/Carleton_University	http://dbpedia.org/resource/Carleton_University
8	computing cloud	null	http://dbpedia.org/resource/Cloud_computing
9	eweek	http://dbpedia.org/resource/EWeek	http://dbpedia.org/resource/EWeek
10	google	http://dbpedia.org/resource/Google	http://dbpedia.org/resource/Google
11	hp	null	http://dbpedia.org/resource/Hewlett-Packard
12	htc corporation	null	http://dbpedia.org/resource/HTC_Corporation
13	huawei	null	http://dbpedia.org/resource/Huawei
14	industry	null	http://dbpedia.org/resource/Industry
15	labs	null	http://dbpedia.org/resource/Bell_Labs
16	microsoft	http://dbpedia.org/resource/Microsoft	http://dbpedia.org/resource/Microsoft
17	motley fool	null	http://dbpedia.org/resource/The_Motley_Fool
18	network world	null	http://dbpedia.org/resource/Network_World
19	ottawa	null	http://dbpedia.org/resource/Ottawa
20	ottawa citizen	null	http://dbpedia.org/resource/Ottawa_Citizen
21	readwriteweb	null	http://dbpedia.org/resource/ReadWriteWeb
22	reports		http://dbpedia.org/resource/Consumer_Reports
23	researcher	null	http://dbpedia.org/resource/Researcher
24	samsung	http://dbpedia.org/resource/Samsung_Group	http://dbpedia.org/resource/Samsung_Group
25			http://dbpedia.org/resource/Samsung_Heavy_Industries
26			http://dbpedia.org/resource/Samsung_Lions
27			http://dbpedia.org/resource/Suwon_Samsung_Bluewings
28	sydney morning	null	http://dbpedia.org/resource/The_Sydney_Morning_Herald
29	telus	http://dbpedia.org/resource/Telus	http://dbpedia.org/resource/Telus
30	uk	null	http://dbpedia.org/resource/United_Kingdom
31	us	null	http://dbpedia.org/resource/United_States

Figure 5.9: Alchemy vs SoSeM. Social Semantic Matching on SocialHead twitter account.

concepts)?

Not surprisingly, hierarchy concepts from the first two levels were best rated by users (6.5/10 on average). This was expected as participants claimed that a categorized user profile would help them have a clear view of their image in the community. However, hierarchy concepts from above were not validated (3/10 average score), which means that users are not willing to have too abstract profiles (i.e. sharing about Twitter does not mean the user is expert in Human Computer Interaction. This result allowed us to set the maximum authorized hierarchical expansion for profile concepts to 2.

With regards to the other dimensions in the expansion, direct expansions were well ranked by users, but not the expansions that include concepts from the same category. However, for the sake of diversity, we still include such concepts in the profile, but set the parameters of the propagation algorithm to high thresholds in the case of such concepts in order to expand only if high expertise is associated.

Facebook: 62.962962962963
Networks: 60.740740740741
Sociology: 57.058823529412
American Jews: 53.492063492063
American billionaires: 44.694862979068
Social systems: 44.444444444444
Student culture: 44.426406926407
Social networks: 43.846120984279
Web 2.0: 42.5
Social sciences: 42.261904761905
Facebook employees: 41.107103422893
Social media: 41.025641025641
Network theory: 36.947712418301
American atheists: 36.899415204678
Cultural economics: 36.842105263158
Systems theory: 35.614379084967
Social network services: 33.477650063857
Value: 33.333333333333
Social psychology: 31.642512077295
Graph theory: 31.632837750485
Living people: 29.65873015873
Community building: 28.800915331808
Harvard University alumni: 27.501780626781
American computer programmers: 24.750196625197

Figure 5.10: Filtered Semantic Expansion of concept Facebook.

In order to further enrich expertise profile another dimension for semantic expansion could be represented by the inclusion of metadata from shared hyperlinks in microposts. Therefore, we performed an additional experimentation in order to understand wheather users accept such content as an extension of their expertise.

The experimentation required the development of an additional component that is capable of extracting metadata from a HTML page. The result of the extracted metadata was includede in the tag cloud representing the user profile and users were asked to rate these concepts.

The extraction of metadata from a HTML page was performed in the following way. The following sources of information can be used to understand the content of a HTML page:

- Metadata extraction from the description of the html page. This can include (i) the title of the document, (ii) the keyword set and (iii) the description text. If a term appears in several parts of the meta description, its occurence is weighted according to the position of the tag.

- Extraction of keywords from the content of the web page. Most online keyword extraction services have such a feature. More concretely, this service counts the number of occurrences of semantically-defined entities (i.e. concepts and in-

stances) that are represented by each term t, when they are identified in the document d. In our case, we employed the corresponding feature of AlchemyAPI.

- The use of external sources where people can manually tag a web page, e.g. Delicious. Such manually entered tags can sometimes give a better description of the content.

The exploration of these different techniques resulted in the following observations:

- AlchemyAPI returned several relevant keywords, including higher-level concepts but also too many meaningless keywords. These keywords were eliminated by the conceptualization algorithms, as no corresponding concept was found in DBPedia.
- The HTML-based metadata extraction mostly emphasized keywords that were found in the title description of the page. Nevertheless, this is not the case on all web pages, therefore this extraction method is not very useful.
- The same applies to Delicious tags: it returns few keywords for a web page, which better describe their topic, as their is a consensus between people who tag the given resource.

Based on this analysis our decision was to consider the content of the HTML page as the primary source of additional concepts to be included in the user profile. The average rating of users for such concepts was of 4/10, showing that such concepts can be sometimes relevant, but do not reflect expertise. However, category concepts computed based on the keywords in the text of the webpage were better rated and therefore such concepts are included in the user profiles vector.

5.2.4 Keyword-based profiles vs. Concept-based profiles

In order to rate the representativity of expertise profiles, we asked each volunteer participant to report on a 4-point Lickert scale their answers to the following question "How well does this tag cloud represent your expertise on average?" with respect to their profile in the case of a (i) keyword-based construction and (ii) conceptual construction. With regards to the conceptual construction, the expansion algorithm was included in this step (except the expansion with concepts from shared URLSs).

We observe that these ratings are quite homogeneous for each participant, despite the heterogeneity of proposed profiles. The average representativity was rated 2.5 in the case of keyword-based profiles and 3.2 in the case of expanded conceptual profiles. This is a significant gain (0.7) showing the usefulness of the semantic expansion and the underlying conceptual analysis. ‘

5.3 Discussion on the Implementation and Evaluation of the Framework

Our results comfort us in the sense that, despite the novelty of this kind of representation of profiles from micropost data, the concept and visualization of profiles is understandable by users.

Another interesting observation with regards to our expertise scoring mechanism is the fact that the inclusion of entropy and sentiment in the scoring function allows to partially limit users that retweet automatically or are not real users.

An interesting observation is also the fact that friends answer questions more quickly, however the answers are generally short. People not yet connected to the user provide much larger answers, but after a longer time period. We consider that this reflects the real nature of social networking behavior and shows the importance of mixing friends and more distant connections in expert recommendation (in the current scenario this is composed of 50 - 50 %. An additional explanation of the social link to a recommended connection would probably further increase the number of interactions and connections. A model and implementation of a component generating such explanasion can be found in [Lajmi 2009].

Our main objective in the evaluation scenarios was to proove the added-value of conceptualizing the knowledge of a social search system. The increase in representativity shows that our approach is promising as it allows to have a rich profile even in the case the user is a newcomer and has few messages. Currently, the Tagging Beak generated 120 new connections between Twitter users, which is also a promising results for our hypothesis, claiming that social search is a better strategy for information retrieval in social platforms. Further evaluations are necessary in order to validate the other dimensions of our approach, i.e. the privacy management.

CHAPTER 6

Conclusion and Perspectives

Contents

In this chapter, we summarize our research, we recall our contributions and our findings. Aiming to address limitations existing in current approaches to explore the content shared in social platforms, and implicitly social search strategies built on top of social platforms, this thesis elaborates on the construction of user profiles from shared posts in these platforms.

In the first part of the thesis, we reviewed the research fields in which this work is framed: i) *Social Network Analysis*, (ii) *Semantic Metadata Management* and (iii) *User Modeling*. Finally, we consider existing approaches for (iv) *Privacy Management*, a cross-field, as this is fundamental in systems that allow content sharing and some kind of social interaction between users. The objective of this first part was to draw clearly the limits of existing work that could answer our challenge and to identify the corresponding contribution areas.

In a follow-up of this part, we review the related work in the recent field of social search. The overall revision of these research fields allow to have an upper-view of recent research directions in semantic-based information representation and retrieval.

In the second part of the work, we presented our framework proposal for implementing social search on top of an existing social platform. We report on several experimentation conducted either to validate individual algorithms, or test more high-level concepts.

In order to achieve this, the following goals have been reached:

- The definition of a formal conceptual model for the representation of the user.

This is achieved by the connection of terms in the user model to concepts in Linked Data knowledge bases.

- The definition and implementation of a toolkit, composed of algorithms that allow to extract knowledge from content productions of the user and associate them to external concepts from the knowledge base.
- The definition and implementation of a social privacy management strategy, that takes into account the different dimensions of users specified in the beginning of the thesis. This approach allows currently to generate an explanation in the form of the most relevant profile attributes for each recommended user with regards to the topic of a question. For each said attribute, a corresponding granular variation is computed based on our approach.
- The building of an online available proof-of-concept social search system, on top of an existing social platform, i.e. Twitter, that implements our social search strategy and that allows the joint evaluation of the above proposals. It is important to mention that this system is currently used by a large community of Twitter users and counts more than 18000 registered users.

In the following, we present the conclusions and summarize the contributions achieved in this research work, and we discuss the limitations of the proposals, along with future research directions to address.

6.1 Summary and Contributions

The final result of this thesis consists of models and algorithms that are integrated and demonstrated in a social search engine available online (`www.tbeak.com`), that allows users to ask questions in natural language and have access to community members who are expert and interactive on the related subject areas.

6.1.1 Contribution to Knowledge Representation in User Models

A term in a message can have several meanings, and the user might be interested in only one of them. Without taking into consideration the meaning of the term, all the items where that term appears could be recommended to the user, whereas only some would be relevant.

The rest of the items would comprise wrong, not useful recommendations (c.f. our contribution to semantic matching). Another reason is the term independence assumption. The fact of an item not having user interest terms explicitly does not necessarily implies that the item is not relevant for the user. Other related terms (by synonymy, hypernymy, hyponymy, etc., relations) could be taken into account to determine the

importance of the item for the user (c.f. our contribution to semantic expansion). This is particularly useful, when the input content is short and unstructured, which is the case of the posts shared in the most common social platforms.

The previous limitations imply that in most of the current user models that are used in social search system, there is a lack of understanding and exploitation of the semantics underlying the user interests. Also, this is the first work addressing this issue in the case of short content to the best of our knowledge.

This is the main reason why we proposed an approach that puts high emphasis on the disambiguation of each item that is about to be injected in the user model. Thus, we identify a first contribution of this thesis, as follows:

The definition of a formal knowledge representation of user preferences extracted implicitly from shared content, which is composed of non-ambiguous concepts from an external knowledge base.

The use of such a conceptual representation of items in user profiles, in contrast to other common approaches based on keywords or items, offers the following benefits:

- Semantic richness. User interests are more accurate, and reduce ambiguity. This enables a better understanding and exploitation of the meanings of items in the user profiles involved in the general social search process.
- Hierarchical representation. The semantic neighborhood of a certain concept can provide additional valuable information about the semantics of the latter.
- Portability. Using ontology-based standard, domain knowledge, item annotations, and user preferences can be easily distributed, adapted or integrated in different systems for different applications. Our review of semantic metadata management models can help developers quickly choose the best model and integrate it into the knowledge layer of the framework.

We emphasize the fact that in the case of social platforms, the use of a conceptual knowledge representation is even more beneficial. More concretely, the semantic expansion of concepts using the knowledge base allows to further approximate user interests. This is necessary, since the short nature of content in such platforms does not motivate users to express all their interests. For example, when sharing about a movie in a blog, a user would probably explain why he/she liked or not the given movie (e.g. there were good actors). However, in a micropost, generally, it remains just a simple opinion, without any explanation.

Therefore, another contribution of our thesis is the design of a novel mechanism that explores Linked Data in order which extends the semantic descriptions of user preferences and through the ontological relations of the involved concepts in Linked Data, called Semantic Expansion. This mechanism allows to overcome several limitations produced by the content sharing habits in such platforms: (i) Sparsity problem.

By applying a semantic expansion, user and item profiles become larger, covering more areas of the conceptual space, and resulting in a higher likelihood of finding user and item similarities and correlations. (ii) Coping with the cold-start problem. The semantic expansion of new user profiles eases their early incorporation and better exploitation in the social search processes. This partly answers a limit of other state-of-the-art social search platforms where users are asked to manually fill their profiles, which often results in only a few very general keywords.

6.1.2 Contribution to Knowledge Extraction from Shared Content

Another significant contribution of our thesis is the overall process of knowledge extraction from shared content in social platforms. The main novelty in the field of research is the fact that we designed a series of algorithms that are specifically tailored to extract the maximum of knowledge possible from short, unstructured microposts. The originality of our toolkit is the projection of knowledge extracted from microposts to a conceptual level, using semantic matching and expansion algorithms and also the scoring of profile concepts, with the sentiment polarity, entropy and statistical sharing patterns, represented by the tf-idf.

6.1.3 Theoretical Contribution to Privacy Management in Social Platforms

The projection of extracted knowledge to a conceptual level allows to completely rethink the way privacy is managed in social platforms and how a user profiles may be used to generate explanation for a recommendation. In existing approaches, as presented in the state of the art, privacy is considered a binary mechanism. A user can either share a piece of information, or not. The privacy management strategy that we introduced on a theoretical level allows going beyond this approach, by considering granular variations of a content artifact. We introduced our approach by considering a criteria that is related to social spheres. The examples afterwards shows that it is easy to adapt the approach also to the case of our proof-of-concept application, where we compute the level of granularity of profile concepts to be shows in a particular conversation space centered on the topic of a question.

6.1.4 Contribution to Social Search Systems: The Tagging Beak

The Tagging Beak is an online available social search system implementing our algorithms for content processing and granular privacy management. Currently it produced promising results, as there are an interesting number of interactions and new connections between people have been created due to the search strategy. Also, this application has had promising feedback from experts in the field of Semantic Web.

6.2 Perspectives

In this work we presented a way of enriching microposts with the help of a semantic knowledge base and build expertise profiles for users of social platforms. All components of this framework can be further improved mainly in the following way:

- The vector-space model responsible for the ranking can be improved with a topic-based model. Also, implementing a similarity measure that takes better into account to semantic links in the knowledge base for the computation of similarities could further show the benefits of manipulating content on a conceptual level in social search scenarios. However, our objective in this work was limited to the knowledge extraction toolkit.
- The current way expertise is computed may be improved with additional statistical measures on the interactivity and influence of the user in the social platform. For this, several additional measures can be performed on the user's community and behavior in the social platform. This may include:
  - The definition of influence in a social network has different definitions. Klout.com, for example, defines influence as "the ability to drive people to action" - "action" might be defined as a reply, a retweet, a comment, or a click.
- Expertise profiles can be further improved by taking into account contextual information with regards to the micropost. In such way, profiles may be clustered according to location, time etc.
- Users should be able to define reachability rules regarding the social sphere who could reach him/her with a particular question. We developed a preliminary model for this issue in [Stan 2009] but this model needs further improvements.
- The framework should allow the integration of multiple social platforms and the analysis layer should perform the aggregation of the social data. Identity disambiguation algorithms could be integrated in order to identify the same user with different logins.

Chapter 7

Appendix

7.1 The Case of Facebook for Social Content Capture and Crawling

7.1.1 Access Authorization to Facebook Social Data

When a Facebook user authorizes a given application, the application gets access to the user's Facebook ID. By default, the application can access all public data in a user's profile, including her name, profile picture, gender, and friends. If the application needs to access other parts of the user's profile that may be private, or if the given application needs to publish content to Facebook on a user's behalf, the application must request extended permissions.

The Facebook crawler needs to access users information at any moment, even if the user is not online. For this purpose, the application needs to be provided with "offlineAccess". This special authorization enables the application to perform authorized requests on behalf of the user at any time. By default, most access tokens expire after a short time period to ensure applications only make requests on behalf of the user when they are actively using the application. The "offlineAccess" special permission makes the access token returned by the Facebook OAuth endpoint long-lived. The application also needs special data permissions from the user to be able to retrieve the data to be analysed [1].

Figure 7.2 shows an example of Facebook authorization.

Facebook Platform supports a number of flows so that the users can be authenticated in web applications via redirects, in JavaScript, or in desktop and mobile applications. Once the user authenticated on Facebook and access authorization accepted, our server retrieves credentials to be able to access user's social data when the user is offline.

Facebook platform seems to be moving to a model where applications must list all the pieces of data they need to access from a user's profile rather than having all that data available automatically. Facebook platform is also moving to an authentication model where all permissions are granted in a single dialog rather than a sequence of many dialogs. These two changes together should increase transparency for users. In our first implementation model, the authorization was given in a sequential mode and

[1] http://developers.facebook.com/docs/authentication/permissions

```
      {
         "id": "9829_141700799187011",
         "from": {
            "name": "Ioana Petrescu",
            "id": "9829"
         },
         "message": "dca airport looks like a crowded parking lot. there is the
longest line ever for airplanes to get to the gates.",
         "link": "http://www.facebook.com/",
         "icon": "http://photos-a.ak.fbcdn.net/photos-ak-
snc1/v27562/151/2254487659/app_2_2254487659_1473.gif",
         "attribution": "Facebook for BlackBerry",
         "actions": [
            {
               "name": "Comment",
               "link": "http://www.facebook.com/9829/posts/141700799187011"
            },
            {
               "name": "Like",
               "link": "http://www.facebook.com/9829/posts/141700799187011"
            }
         ],
         "type": "link",
         "created_time": "2010-07-30T01:14:11+0000",
         "updated_time": "2010-07-30T04:29:27+0000",
         "comments": {
            "data": [
               {
                  "id": "9829_141700799187011_1137594",
                  "from": {
                     "name": "Ron Lai",
                     "id": "2203233"
                  },
                  "message": "hrm.  why?  i am dc bound monday/tuesday.  guess you
won't be around?",
                  "created_time": "2010-07-30T04:29:27+0000"
               }
            ],
            "count": 1
         }
      }
```


Figure 7.1: Example Facebook profile feed in JSON format

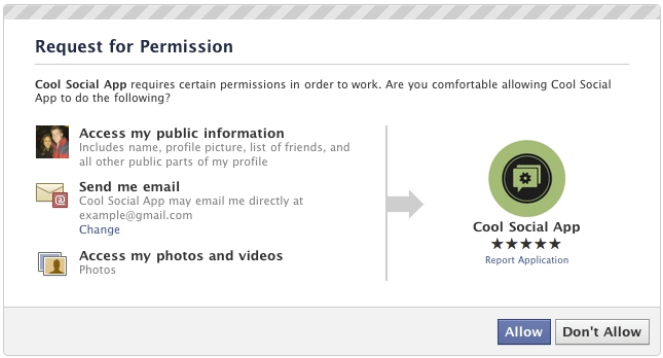


Figure 7.2: Giving permissions to an application

user's information the application was able to access was not listed explicitly in the authorization form.

7.1.2 Data Access in Offline Mode

Facebook Platform is transitioning from using their own authentication system to OAuth 2.0 protocol, an open standard Facebook co-authored with a number of other companies, including Yahoo, Google and Twitter. All of Facebook existing APIs now support OAuth, including the old REST API, and will continue to support the Facebook old authorization scheme as well. OAuth 2.0 is a simpler version of OAuth that leverages SSL for API communication instead of relying on complex URL signature schemes and token exchanges. At a high level, using OAuth 2.0 entails getting an access token for a Facebook user via a redirect to Facebook (see previous description). After the application obtains the access token for a user, the application can perform authorized requests on behalf of that user by including the access token in the Graph API requests.

Once the access credentials retrieved and the special authorizations accorded to the application, the user's social data can be accessed "offline" by using:

- Facebook "Old" REST API. The old REST API is the previous version of the Graph API. It enables an application to interact with Facebook web site programmatically via simple HTTP requests. This API supports several access methods:

  1. Custom authorization signature scheme
  2. OAuth 2.0[2]

- Since April 2010, the new Graph API is recommended to be used. The Graph API enables applications to read and write objects and connections in the Facebook social graph. The Graph API is using OAuth 2.0 for authentication purposes.

The first implementations of the Facebook crawler have used the Facebook "Old" REST API with a custom authorization signature scheme. The Java client used in the first implementation of the crawler was the "Facebook Java API", an open-source project built around what was previously the official Facebook Java client. However this client has not been maintained therefore it does not seem to support the Graph API. Next implementations for the Facebook crawler will use the Graph API with an OAuth2.0 access. Clients like TinyFBGraphClient or BatchFB are good candidates for the new implementation of the crawler using Graph API and OAuth2.0.

[2]The OAuth 2.0 Protocol, draft-ietf-oauth-v2-10 (http://tools.ietf.org/html/draft-ietf-oauth-v2-10, visited the 2010/07/28)

Facebook provides mechanisms which allow clients to combine several requests into a single call. BatchFB library let the client make numerous requests and queries in a natural syntax, then automatically optimize them down to the minimum number of Facebook calls. This library handle aspects related to multiquery therefore it is a serious candidate to be integrated in the new version of the crawler.

7.1.3 Data Retrieved

All elements retrieved by using the "Graph API" are JSON objects and all objects in Facebook can be accessed in the same way. Also all of the objects in the Facebook social graph are connected to each other via relationships. Each element has a unique ID and is related to the users having published the element (in the "from" field). Example of objects retrieved by using the Graph API:

- Friends: https://graph.facebook.com/me/friends
- News feed: https://graph.facebook.com/me/home
- Profile feed (Wall): https://graph.facebook.com/me/feed
- Likes: https://graph.facebook.com/me/likes
- Movies: https://graph.facebook.com/me/movies
- Books: https://graph.facebook.com/me/books
- Notes: https://graph.facebook.com/me/notes
- Photo Tags: https://graph.facebook.com/me/photos
- Photo Albums: https://graph.facebook.com/me/albums
- Videos: https://graph.facebook.com/me/videos
- Events: https://graph.facebook.com/me/events
- Groups: https://graph.facebook.com/me/groups

The introduction to the Graph API[3] is presenting several examples of use. By accessing the links listed in this introduction (user has to be connected in Facebook) data retrieved from Facebook for the connected user is displayed just as the user's data was accessed "offline" by the application by performing HTTPS calls. Particularly interesting for our data analyze is the information figuring on the user's Wall and on the News Feed. For example by accessing the "Profile feed" object, the user's "wall" can

```
{
   "id": "9829_141700799187011",
   "from": {
      "name": "Ioana Petrescu",
      "id": "9829"
   },
   "message": "dca airport looks like a crowded parking lot. there is the
longest line ever for airplanes to get to the gates.",
   "link": "http://www.facebook.com/",
   "icon": "http://photos-a.ak.fbcdn.net/photos-ak-
snc1/v27562/151/2254487659/app_2_2254487659_1473.gif",
   "attribution": "Facebook for BlackBerry",
   "actions": [
      {
         "name": "Comment",
         "link": "http://www.facebook.com/9829/posts/141700799187011"
      },
      {
         "name": "Like",
         "link": "http://www.facebook.com/9829/posts/141700799187011"
      }
   ],
   "type": "link",
   "created_time": "2010-07-30T01:14:11+0000",
   "updated_time": "2010-07-30T04:29:27+0000",
   "comments": {
      "data": [
         {
            "id": "9829_141700799187011_1137594",
            "from": {
               "name": "Ron Lai",
               "id": "2203233"
            },
            "message": "hrm.  why?  i am dc bound monday/tuesday.  guess you
won't be around?",
            "created_time": "2010-07-30T04:29:27+0000"
         }
      ],
      "count": 1
   }
}
```


Figure 7.3: Example Facebook news feed in JSON format

be retrieved. It represents the content a user shares with friends. Next image (Figure 7.3 illustrates the type of object retrieved by accessing the "News feed":

The content of the message as well as the comments related to the message are retrieved and analyzed by the crawler. By default, most object properties are returned when the application makes a query. The application can also choose the fields (or connections) needed by using the "fields" query parameter. For example, this URL will only return the id, name, and picture of the user: https://graph.facebook.com/bgolub?fields=id,name,picture It should be noted that content retrieved from Facebook is subject to Facebook data policies [4].

7.2 The Case of OpenSocial-based systems for Social Content Capture and Crawling

Netlog is an example of a social networking site that can be accessed through the OpenSocial. As its name suggests, OpenSocial is an API that permits applications to access social data in social networking sites in an non-proprietary manner, such that the same code can work with different sites. In practice, the configuration is fairly specific and code needs to be rewritten or "massaged" to work with different social networking sites. The OpenSocial API is a competitor to the Facebook API, with much of the same functionality.

7.2.1 Available data and permissions

- People: information about a user, distinction between a profile owner and a profile viewer (person who look at others' profiles). Relationships: ability to retrieve the friends of a profile viewer or owner.

- Activities: exposition of actions a user has taken in the context of a given container, like profile update, installation of a new social gadget, new high score in a game, gift offered to one friend, etc.

- Messaging: defines ways for reading, posting, and deleting messages between users in the network. OpenSocial defines a number of message types including public messages (such as profile comments), and private messages (messages restricted to certain individuals and groups).

[3] http://developers.facebook.com/docs/api (visited the 2010/07/28)
[4] http://developers.facebook.com/policy/data (visited the 2010/07/28)

7.2.2 Obtaining User Credentials

Communication between the Netlog crawler and Netlog uses the same Oauth mechanism as many other sites. Unlike most other social networking platforms, Netlog validates every third-party application before permitting them access to their social data. This makes it difficult for a third party developer to enter into the Netlog social ecosystem, but ensures that the third party and Netlog can cooperate to their mutual benefit. As a result, the restrictions around privacy aren't as strong and applications have more access to users' data. To crawl Netlog data, a new consumer key and secret must be requested from their developer's program.

7.3 Panorama of Social Platforms

In this section, a description of the most popular social platforms will be given. Each platforrm will be reviewed from the viewpoint of research and technical background, mobility, statistics and its relation to our framework, as well as the business model.

7.3.1 Facebook - `www.facebook.com`

The most used social networking site on the web and often considered the "standard" for social networking services. The majority of social applications seem to be games, especially quizzes and point-and-click time wasters.

From the point of view of research, Facebook is present in many scientific work (312 000 results on Google Scholar, likely due to it being one of the most active social networking platforms (and therefore a source of the most relevant data). Facebook is more concerned with providing a complete commercial platform and leaving innovative research to their partners. It seems likely that most of their internal research has to do with scalability of providing their massive services than innovating on the social aspect.

From a technical viewpoint, Facebook is a customized LAMP stack. API built on PHP, including Facebook Connect for third party sites. It provides single-sign on Facebook authentication API for websites and mobile web applications.

Facebook provides a complete toolkit for mobility, with as there is a dedicated Website for mobile: `http://m.facebook.com/` and the platform is integrated to some mobile phones out of the box (iPhone app, Android 2.0 including HTC Hero).

In conclusion, Facebook is a mandatory source of social information and its integration to the social search framework would further increase information about user's profiles.

7.3.2 Twitter - `www.twitter.com`

Ultra-simple micro-blogging service that has captured a huge amount of media, user and developer attention recently. Users publish a stream of very short messages (140 characters), on every conceivable topic at all hours of the day. Mostly public, although direct messages (between two users) are possible, and publishing may be restricted to "friends". Uncontrolled list of followers for a user's stream. Followers aren't reciprocal. Twitter has supplied an API from the beginning, promoting it as an API for third party devices from the beginning (cars that tweet, scales that tweet, Tower bridge tweets, etc.).

Twitter's Web interface uses Ruby on Rails and the Backend is written in Scala. The API is separated into two: REST API and Search API (to be consolidated in the future): get message streams, create messages, search for topics, examine social ties (friends and followers), change profile, favorites.

From the point of view of mobility, Twitter is inked to telecom networks through SMS in various countries (but not France). Almost every mobile OS has multiple Twitter clients written for it. The API permits searching by location (longitude/latitude).

Interestingly, according to the site, Twitter has no specific business model for the moment. They are looking to create a community first, however they collect personally identifiable information about its users to be shared with third parties and partners, and retain the right to sell it. However, Twitter has several partners: Google and Microsoft Bing to index public tweets. LinkedIn to share status updates. Orange for updates directly from mobile/text messages/MMS (including photos).

With regards to our framework, Twitter is the most interesting platform, as source of social information (micro-communications for users, can be easily captured and analyzed), users following other users. Very important site to follow and interoperate for reasons of "hype".

7.3.3 LinkedIn - `www.linkedin.com`

LinkedIn is a platform for professional networking, so profiles and messages are encouraged to be professionally related (like an online resume as opposed to a diary). The first principle is that people can't directly send messages or notifications to other members without a relationship, but can ask for introductions via their chain of acquaintances. An interesting service of LinkedIn is its capability to automatically compile statistics about participating companies (such as the male-to-female ratio) to help job searches, social search for specific professional questions (LinkedIn Answers). Discussion groups (LinkedIn Groups) for companies, alumni, professional development and interests are also supported.

Often described as "Facebook in a suit", LinkedIn focuses on the professional aspects of social networking. People that are too nervous about revealing their personal exchanges on other websites (such as Facebook) often exchange LinkedIn contacts eas-

ily. In this sphere, they've made an effort to keep a serious image - applications tend to be professionally-oriented. The site's design tends to discourage frivolous interactions in order to keep their distinguished appeal (to the audience that doesn't want to share Facebook-like interactions in a professional environment).

The system can be interesting for the framework of social search, if planned to exploit it in a professional context.

7.3.4 MySpace - `www.myspace.com`

As top competitor with Facebook, very similar features: my profile, my friends, browsing friends of friends, wall (status updates), pokes, shared media, photos, videos, discussion groups, forums, events (with configurable access control). The special feature is the availability of special profiles for musicians and producers, emphasis on music fandom and sharing your original songs.

The main competitor to Facebook, it used to be the most visited social networking site, although this figure was contested because its design forced many page views. As a competitor, they include some sort of parallel to almost every feature of Facebook, and try to address where Facebook lacks, specifically: (i) based on OpenSocial instead of proprietary API, (ii) investigation into the feasibility of Data Portability. They've found their niche in music promotion. MySpace has a reputation for embracing ugly UI design and is considered more popularist in comparison to Facebook (soldiers use MySpace, officers use Facebook).

MySpace Developer Platform (MDP) API built on OpenSocial and a OAuth authentication API for websites and mobile web applications. MySpaceId is a competitor to Facebook Connect, allowing third parties to use MySpace credentials to sign onto their website.

This website can provide mandatory source of social information and bridge via an OpenSocial application ideally customized for MySpace.

7.3.5 Netlog - `www.netlog.com`

Very, very similar in function and vision to Facebook (and initially even called Facebox). It has captured a large European youth market, likely because it entered the European market early with a competent internationalized website. Its internationalization support is still a key differentiator, along with its commitment to open APIs and standards for interoperability with other social networks and applications. Notably, it is based on an OpenSocial container with extensions.

It's technical background is represented by an Application API is OpenSocial with Netlog extensions to help access their specific functionality. The website feels much slower than Facebook. Netlog has 59M members (80% are between 14 and 25 years old),

150M visits/month. No vision on unique visits/month. Page view market leader in Belgium, Italy, Austria, Switzerland, Romania, Turkey. Second place in the Netherlands, Germany, France, Portugal. 2008 Website of the Year by MetrixLabs (independent research firm), 2008 People's Choice from Mashable!

7.3.6 Delicious - `http://www.delicious.com`

Social bookmarking service that categorizes users' preferred web sites by non-hierarchical tags/folksonomy.

The application saves your bookmarks online, meaning that they are available from any computer (home and work), and can be made public or private. The social aspect comes from the shared bookmarks. Delicious has "networks" which are people you know (i.e. friends) and "fans" who subscribe to other users that have interesting or relevant collections of links.

One of the top cited sites for tagging, an extremely good data set for discovering emerging folksonomy and behavior from public tagging actions. It provides a Simple REST-like API and RSS feeds for aggregation. A user can download his or her own data through the site's API in an XML or JSON format (including networks and fans). There is a Firefox plugin used for persisting bookmarks across computers.

This system can provide a very interesting dataset for the framework in the form of source of social information (networks and fans for explicit links between people). Very important source of crowd-sourced categorization of web pages in general. Important source of user interests declared publicly, for profiling purposes.

7.3.7 StumbleUpon - `http://www.stumbleupon.com`

Just like del.icio.us CiteULike or Diigo, stumbleupon is social bookmarking website that allows users to rate, tag and share comments about websites, photo, video. StumbleUpon recommends on this basis tagged and rated documents to its subscribers with similar profiles.

Stumbleupon provides a browser (FF, IE) extension to install as a browser toolbar that allows people to tag or rate web elements. Users can also "stumble" web elements through the StumbleUpon website. The StumbleUpon model of web browsing resembles "channel surfing" (zapping). In brief, every time the user triggers the service, they are automatically directed to any page that the system feels they would find interesting, based on their history and the history of previous users. The user can then add a comment and rate a page (thumbs up/thumbs down). Users can also form a social network by linking to their friends using the service, in order to recommend "socially endorsed pages".

StumbleUpon gives access to open API called Su.pr. Su.pr provides real time analytics for all of generated links as well as the ability to post your Su.pr links to other

social media services such as Twitter and Facebook.

It could act as a data source for the framework, especially with user-generated profile information. Currently there is no open API to fetch this data.

7.3.8 Foursquare - `http://www.foursquare.com`

Foursquare is a location based social network that incorporates gaming elements. Android and iPhone application available on Android Market and Apple Store. The main idea is that people can check in their current location and see what is interesting nearby, i.e. recommendations from others. In addition, users can publish their location to their friends (pings, or push notifications).

Foursquare is a kind of geo-location based social game, where people get points and badges for checking into places. The person who checks into a place the most becomes the "Mayor." Thus, different places like restaurants and bars offer promotions to mayors in order to attract them. Promos are presented as actual ads in the application.

It works with an API based on HTTP for checking in, getting profile details, friends, venues, tips. Android client is open source.

This platform is not really related to our scope, although it could be a potentially interesting source of geolocal social information.

7.3.9 Aardvark - `http://www.vark.com`

Social Search Engine: the main idea is to find experts in a subject, not web pages. More concretely, Vark is a social service that allows users to get quick, quality answers to questions from their extended social network. The user asks questions via an instant messenger or email and the question is propagated to their contacts and their respective contacts.

The user can manually update their profile to better reflect their expertise, and they can manually tag the question if the automatic system is off. They can refer questions to people in their social circles, and they can share answers and conversations.

7.3.10 Swotti - `http://www.swotti.com`

Swotti is a platform to search for opinions on a product from the Internet, extracting opinions (like, dislike etc.) and giving a ranking to products directly from them. Depending on the topic, the tool can take more then one minute before giving a result, and there are many products for which the platform does not find any result.

The platform is not able to give results on a targeted product. The best result we could obtain concerned a trademark. There is no interaction with the social networks of the user.

The platform is related to BuzzTrend company. Swotti is a semantic search engine that helps consumers to gather product reviews, opinions, and articles. The site indexes over 3 million reviews, good and bad. After searching, Swotti returns a number of visual results that allow the user to easily evaluate the collective opinion on their product in question.

Despite being a "semantic search engine", Swotti performs only simple, rating-based opinion mining. Thus, there is no semantic analysis for the retrieval of opinions. This is a problem, as products with similar names exist and in this case the system retrieves an inaccurate aggregation.

7.3.11 Posterous - `http://www.posterous.com`

Posterous is a micro-blogging platform that focuses on publishing your content and life, and providing a central place to look for your updates on many social networking sites. It can automatically update these sites as well, including social networking sites, other blogs, and other media publishing sites.

Themed home pages which are easily redirectable to custom domain names, and group blogs with multiple authors. Users post messages on Posterous by sending an email to post@posterous.com. Otherwise, messages once posted seem to be visible by everyone.

API is designed to add posts and get feeds using posterous or Twitter credentials. Posterous integrates with many sites:

- Social networking: Facebook (and Facebook Page), Twitter, LinkiedIn, FriendFeed, Jaiku, Plurk, Identica.
- blogs: Blogger, Tumblr, Livejournal, Shopify, Typepad, Xanga, Wordpress, MovableType, Drupal, and other blogs that support the MetaWebLog API
- media: Flickr, Picasa, YouTube, Vimeo, Google Video, Viddler, Blip.tv, Scribd
- Delicious.

It is mainly designed to be used in a mobile environment, for "lifestreaming" and it is especially useful for updating by email, including attachments as photos, documents or videos. The PicPosterous iPhone application integrates with iPhone photo albums.

It is a competitor for aggregation with our framework (especially on the publishing end, synchronizing all the statuses of users' social sites). Could be an interesting pre-aggregated source of social information, although thereis currently no API available to get the pre-aggregated social information.

7.3.12 Silentale - `http://www.silentale.com`

Silentale is a social network aggregator, including users' contact lists, social networking websites, and mailboxes. The user defines the different communication sources to be aggregated and the tool ensures the combination of the contacts, conversations, etc. and keeps track of who knows who and which conversation has been sent from whom.

Silentale's aggregation of social networks is very basic; it combines different specified sources in a centralized location while retaining the origin of the communication, supporting identity disambiguation, archiving, and export of the conversations of a friendly form. Silentale offers and relies on five main services:

- (i) Connectors: an information channel which is connected to a source of data for Silentale. These sources may include address books, emails, social networks or SMS accounts, etc. Concrete examples are Facebook Friends, Google Contacts, most email accounts (IMAP / POP3), Gmail, AOL Mail, Yahoo! Mail, Hotmail/Live Mail, Twitter.

- (ii) People Books: an address book that supports contact disambiguation, and presents the contacts in a unified way no matter the source they come from. Every time a user wants to communicate, the system displays the different possible communication platforms and brings together all their contact details: emails, phone numbers, addresses, social profiles and screen-names.

- (iii) Timeline: an interesting display of the communication history of all the conversations ranked in chronological order. (iv) Conversation manager: a channel made to rebuild the discussion/conversation held between people. This is similar to what Google does in Gmail.

- (v) Archive: an archiving utility for messages and conversations whether they are emails, SMS, tweets, etc., safeguarding them from any accidental loss or deletion.

There is no big research investment at this stage in this platform. There are however some interesting and innovative services like the Time Traveler service which enables users to travel through conversations. The technical background is to simply and continuously capturing one message at time whenever it is happening. API for third party application is available (push, pull % database) (e.g. Android application for SMS). The tool is able to find ALL the identities of the contacts of a person. From a technical perspective, this is possible since the system has the different identities of the users on the different platforms (i.e., communication means).

7.4 Overview of Popular Online Social Tagging Services

In the next sections below we identify currently available online tools for automated tagging and provide a brief evalution of each of them based on a set of criteria.

It should be noted that there is an increasing number of Web services available that perform named entity recognition on textual documents via a Web interface. The majority and the best of such services are not open source or free. There are some free web services (e.g.,tagthe.net `http://www.tagthe.net/`) but they generally provide poor quality performance. Although the majority of services are commercial, some also have free components/versions with limited functionality/usage (e.g., 10,000 requests/day). Most popular services that apply this kind of restriction include:

- Evri
- OpenCalais
- AlchemyAPI

There are also many web services that to all intents and purposes are commercial because the amount of permitted free usage is very small: Meaningtool, Complexity Intelligenece, TextDigger.

7.4.1 OpenCalais - `http://www.opencalais.com`

- OpenCalais is a product of Thomson Reuters that provides an open API that has been widely adopted by the open source community.
- Identifies specific entities, events and relations from the web and news domain (e.g., company merger, natural disaster, product recall, conviction etc). Also suggests social tags.
- A full list of available entities is available here: `http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions`
- Many entities are identified using Calais URIs, some sameAs links to DBPedia and Freebase
- Supports disambiguation of companies, geographical locations and electronics products
- Results available as: RDF/XML, Microformats, custom XML (Simple Format), JSON:
- Provides character offsets that can be used to insert tags into content

- Free for up to 50,000 requests per day after registering for API key, subscription plans above that. Works on documents up to 100K.
- Supports English, French and Spanish
- Detailed documentation available on the website including RDF schema and demo
- It is the semantic tagging engine behind the OpenPublish platform (integrated with Drupal and WordPress)
- ClearForest `http://www.clearforest.com/` - also have a commercial product called OneCalais

7.4.2 AlchemyAPI - `http://www.alchemyAPI.com`

- Automatically tags web pages, textual documents, scanned document images. Supports OCR to analyse scans of newspapers, documents etc.
- Supports multiple languages (English, Spanish, German, Russian, Italian + others)
- Named Entity Extraction API identifies specific entities including people, companies, organisations, cities, geographic features, anniversaries, awards, holidays etc.
- Entities identified by URIs from Linked Open Data (LOD) sources e.g. DBPedia, Freebase, UMBEL, CIA Factbook
- Disambiguation support (although seems to be missing disambiguated URIs for "person" entities)
- Formats: XML, JSON, RDF, Microformats
- Requires an access key to access the API
- Free for up to 30,000 calls per day, can pay for commercial support.
- Detailed documentation available on website including RDF schema and online demo http://www.alchemyapi.com/api/entity/

7.4.3 Zemanta - `http://www.zemanta.com`

- Identifies the following entities: persons, books, music, movies, locations, stocks, companies (documentation does not mention events).
- Also returns related tags, categories, pictures and articles.

- Free for up to 10,000 API calls per day. Subscription plans above that.
- Returns RDF/XML, JSON, or custom XML
- Documentation says it supports custom taxonomies

7.4.4 TagTheNet - `http://www.tagthe.net`

- Returns custom XML containing tags identifying topics, locations, persons, (but not e Also tags for title, size, content-type, author and language of the source document.
- Does not markup content (or indicate location of entities within content).
- Tags are text only (no identifiers or ontology)
- Uses statistical approach (from FAQ). Analysis component is written in Java.
- Free to use as-is. No limitations on use but also no service level guarantees.
- Can invoke via HTTP requests

7.5 Semantic Community Navigation in the Tagging Beak

Besides its main role, the implementation of social saerch on an existing social platform, the Tagging Beak offers users means to better navigate in their community and to discover people who might be interesting for their information needs. This is achieved with the graph metaphor, where nodes are people and concepts and links represent the fact that a concept was shared by a user. In this way, users have a general view of their community and can quickly identify interesting people or communities 7.4.

Figure 7.4: Upper-View of the Tagging Beak: semantic community navigation interface

Bibliography

[311 2006] Harvesting social knowledge from folksonomies, volume Novel systems and models, 2006. (Cited on page 39.)

[Abel 2008a] Fabian Abel. *The benefit of additional semantics in folksonomy systems.* pages 49–56, 2008. (Cited on pages 40 and 62.)

[Abel 2008b] Fabian Abel, Nicola Henze and Daniel Krause. *Ranking in folksonomy systems: can context help?* In CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, pages 1429–1430, New York, NY, USA, 2008. ACM. (Cited on page 42.)

[AdaptiveBlue 2009] Faviki Freebase Yahoo! Zemanta AdaptiveBlue DERI (NUI Galway) and Zigtag. *CommonTag: Open Tagging Format* . 2009. [Online; accessed 20-October-2010]. (Cited on page 61.)

[Amer-Yahia 2009a] Sihem Amer-Yahia, Jian Huang and Cong Yu. *Building community-centric information exploration applications on social content sites.* In SIGMOD Conference, pages 947–952, 2009. (Cited on page 3.)

[Amer-Yahia 2009b] Sihem Amer-Yahia, Laks V. S. Lakshmanan and Cong Yu. *SocialScope: Enabling Information Discovery on Social Content Sites.* CoRR, vol. abs/0909.2058, 2009. informal publication. (Cited on page 76.)

[Angeletou 2008] Sofia Angeletou, Marta Sabou and Enrico Motta. *Semantically enriching folksonomies with FLOR.* In Proceedings of the CISWeb Workshop, located at the 5th European Semantic Web Conference ESWC 2008, 2008. (Cited on pages 48 and 80.)

[Baldauf 2007] Matthias Baldauf, Schahram Dustdar and Florian Rosenberg. *A survey on context-aware systems.* Int. J. Ad Hoc Ubiquitous Comput., vol. 2, pages 263–277, June 2007. (Cited on page 67.)

[Bao 2007] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei and Zhong Su. *Optimizing web search using social annotations.* In Proceedings of the 16th international conference on World Wide Web, WWW '07, pages 501–510, New York, NY, USA, 2007. ACM. (Cited on pages 4 and 42.)

[Barabási 1999] Albert László Barabási and Réka Albert. *Emergence of Scaling in Random Networks.* Science, vol. 286, no. 5439, pages 509–512, 1999. (Cited on page 35.)

[Barker 2009] Ken Barker, Mina Askari, Mishtu Banerjee, Kambiz Ghazinour, Brenan Mackas, Maryam Majedi, Sampson Pun and Adepele Williams. *A Data Privacy Taxonomy.* In BNCOD, pages 42–54, 2009. (Cited on pages 8, 67, 72, 81 and 126.)

[Berners-Lee 2001] Tim Berners-Lee, James Hendler and Ora Lassila. *The Semantic Web.* Scientific American, pages 34–43, May 2001. (Cited on page 50.)

[Besmer 2009] Andrew Besmer and Heather Richter Lipford. *Tagged photos: concerns, perceptions, and protections.* In Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson and Saul Greenberg, editeurs, CHI Extended Abstracts, pages 4585–4590. ACM, 2009. (Cited on page 73.)

[Binder 2009a] Jens Binder, Andrew Howes and Alistair Sutcliffe. *The problem of conflicting social spheres: effects of network structure on experienced tension in social network sites.* In Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson and Saul Greenberg, editeurs, CHI, pages 965–974. ACM, 2009. (Cited on page 33.)

[Binder 2009b] Jens Binder, Andrew Howes and Alistair Sutcliffe. *The problem of conflicting social spheres: effects of network structure on experienced tension in social network sites.* In Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson and Saul Greenberg, editeurs, CHI, pages 965–974. ACM, 2009. (Cited on page 74.)

[Binder 2009c] Jens Binder, Andrew Howes and Alistair Sutcliffe. *The problem of conflicting social spheres: effects of network structure on experienced tension in social network sites.* In Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson and Saul Greenberg, editeurs, CHI, pages 965–974. ACM, 2009. (Cited on page 74.)

[Blanchard 1998] A. Blanchard and T. Horan. *Virtual communities and social capital.* Social Science Computer Review, vol. 16, no. 3, page 293, 1998. (Cited on page 17.)

[Borgatti 2006] Stephen P. Borgatti. *Identifying sets of key players in a social network.* Computational and Mathematical Organization Theory, vol. 12, no. 1, pages 21–34, 2006. (Cited on page 36.)

[Breslin 2005] John G. Breslin, Andreas Harth, Uldis Bojars and Stefan Decker. *Towards Semantically-Interlinked Online Communities.* The Semantic Web: Research and Applications, pages 500–514, 2005. (Cited on page 62.)

[Brickley 2004] Dan Brickley and R. V. Guha. *RDF Vocabulary Description Language 1.0: RDF Schema.* W3C Recommendation, feb 2004. `http://www.w3.org/TR/rdf-schema/`. (Cited on page 50.)

[Brin 1998] Sergey Brin and Lawrence Page. *The Anatomy of a Large-Scale Hypertextual Web Search Engine.* Computer Networks and ISDN Systems, vol. 30, no. 1-7, pages 107–117, April 1998. (Cited on page 41.)

[Broekstra 2002] Jeen Broekstra, Arjohn Kampman and Frank van Harmelen. *Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema.* In The Semantic Web – ISWC 2002: First International Semantic Web Conference Sardinia, Italy, pages 54–68. 2002. (Cited on page 75.)

[Brooks 2006] Christopher H. Brooks and Nancy Montanez. *Improved annotation of the blogosphere via autotagging and hierarchical clustering.* In Proceedings of the 15th international conference on World Wide Web (WWW2006, pages 625–632, Edinburgh, Scotland, 2006. (Cited on pages 42 and 80.)

[Budanitsky 2006] Alexander Budanitsky and Graeme Hirst. *Evaluating WordNet-based Measures of Semantic Distance.* Computational Linguistics, vol. 32, no. 1, pages 13–47, 2006. (Cited on page 43.)

[Budura 2009] Adriana Budura, Daniela Bourges-Waldegg and James Riordan. *Deriving Expertise Profiles from Tags.* In CSE '09: Proceedings of the 2009 International Conference on Computational Science and Engineering, pages 34–41, Washington, DC, USA, 2009. IEEE Computer Society. (Cited on page 71.)

[Burke 2009] Moira Burke, Cameron Marlow and Thomas Lento. *Feed me: motivating newcomer contribution in social network sites.* In CHI '09: Proceedings of the 27th international conference on Human factors in computing systems, pages 945–954, New York, NY, USA, 2009. ACM. (Cited on page 29.)

[Camarinha-Matos 2004] Luis M. Camarinha-Matos and Hamideh Afsarmanesh. *A multi-agent based infrastructure to support virtual communities in elderly care.* IJNVO, vol. 2, no. 3, pages 246–266, 2004. (Cited on page 15.)

[Cantador 2008] Ivan Cantador, Martin Szomszor, Harith Alani, Miriam Fernandez and Pablo Castells. *Enriching Ontological User Profiles with Tagging History for Multi-Domain Recommendations.* In 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008), June 2008. (Cited on pages 48 and 80.)

[Carmagnola 2007] Francesca Carmagnola, Federica Cena, Omar Cortassa, Cristina Gena and Ilaria Torre. *Towards a Tag-Based User Model: How Can User*

Model Benefit from Tags? In Cristina Conati, Kathleen F. McCoy and Georgios Paliouras, editeurs, User Modeling, volume 4511 of *Lecture Notes in Computer Science*, pages 445–449. Springer, 2007. (Cited on pages 57 and 69.)

[Carminati 2009] Barbara Carminati, Elena Ferrari, Raymond Heatherly, Murat Kantarcioglu and Bhavani M. Thuraisingham. *A semantic web based framework for social network access control.* In Barbara Carminati and James Joshi, editeurs, SACMAT, pages 177–186. ACM, 2009. (Cited on pages 73 and 81.)

[Cattuto 2008] Ciro Cattuto, Dominik Benz, Andreas Hotho and Gerd Stumme. *Semantic Analysis of Tag Similarity Measures in Collaborative Tagging Systems*, May 2008. (Cited on page 43.)

[Cohen 1987] Paul R. Cohen and Rick Kjeldsen. *Information retrieval by constrained spreading activation in semantic networks.* Inf. Process. Manage., vol. 23, no. 4, pages 255–268, 1987. (Cited on pages 71 and 115.)

[Consortium 1995] Versit Consortium. *Electronic Business Card Spacification.* `http://www.w3c.org/Submission/vcard-rdf/`, 1995. [Online; accessed 20-October-2010]. (Cited on page 56.)

[Constant 1994] David Constant, Sara B. Kiesler and Lee S. Sproull. *What's Mine Is Ours, or Is It? A Study of Attitudes about Information Sharing.* Information Systems Research, vol. 5, no. 4, pages 400–421, 1994. (Cited on page 17.)

[Crestani 2000] Fabio Crestani and Puay Leng Lee. *Searching the web by constrained spreading activation.* Inf. Process. Manage., vol. 36, no. 4, pages 585–605, 2000. (Cited on pages 71, 81 and 115.)

[Crofts 2006] Nick Crofts, Martin Doerr, Tony Gill, Stephen Stead and Matthew Stiff, editeurs. Definition of the CIDOC conceptual reference model. ICOM/CIDOC CRM Special Interest Group, October 2006. (Cited on page 49.)

[Culotta 2004] Aron Culotta, Ron Bekkerman and Andrew Mccallum. *Extracting Social Networks and Contact Information from Email and the Web.* In In CEAS-1, 2004. (Cited on page 37.)

[Dasgupta 2008] Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan, Dipanjan Chakraborty, Sougata Mukherjea, Amit Anil Nanavati and Anupam Joshi. *Social ties and their relevance to churn in mobile telecom networks.* In Alfons Kemper, editeur, EDBT, volume 261 of *ACM International Conference Proceeding Series*, pages 668–677. ACM, 2008. (Cited on page 36.)

[del Castillo 2009] Joan del Castillo and Jalila Daoudi. *Estimation of the generalized Pareto distribution.* Statistics and Probability Letters, vol. 79, no. 5, pages 684 – 688, 2009. (Cited on page 35.)

[Domingos 2005] P. Domingos. *Mining Social Networks for Viral Marketing.* IEEE Intelligent Systems, vol. 20, no. 1, pages 80–82, 2005. (Cited on page 37.)

[Echarte 2007] Francisco Echarte, José Javier Astrain, Alberto Córdoba and Jesús E. Villadangos. *Ontology of Folksonomy: A New Modelling Method.* In Siegfried Handschuh, Nigel Collier, Tudor Groza, Rose Dieng, Michael Sintek and Anita de Waard, editeurs, SAAKM, volume 289 of *CEUR Workshop Proceedings.* CEUR-WS.org, 2007. (Cited on page 62.)

[Esuli 2006a] Andrea Esuli and Fabrizio Sebastiani. *Determining Term Subjectivity and Term Orientation for Opinion Mining.* In EACL. The Association for Computer Linguistics, 2006. (Cited on page 50.)

[Esuli 2006b] Andrea Esuli and Fabrizio Sebastiani. *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining.* In Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation, Genova, IT, 2006. (Cited on page 122.)

[Fitzpatrick 2005] Brad Fitzpatrick. *OpenID Decentralized Authentification Initiative.* `http://openid.net/`, 2005. [Online; accessed 20-October-2010]. (Cited on page 56.)

[Garcia-Silva 2011] Andres Garcia-Silva, Oscar Corcho, Harith Alani and Asuncion Gomez-Perez. *Review of the state of the art: Discovering and Associating Semantics to Tags in Folksonomies.* Knowledge Engineering Review, vol. 26, no. 4, December 2011. (Cited on page 54.)

[Garcia 2009] Andres Garcia, Martin Szomszor, Harith Alani and Oscar Corcho. *Preliminary Results in Tag Disambiguation using DBpedia.* September 2009. (Cited on page 49.)

[Gauch 2002] Susan Gauch, Jason Chaffee and Alexander Pretschner. *Ontology-Based User Profiles for Search and Browsing*, 2002. (Cited on page 55.)

[Gemmell 2008] J. Gemmell, A. Shepitsen, M. Mobasher and R. Burke. *Personalization in Folksonomies Based on Tag Clustering.* In Proceedings of the 6th Workshop on Intelligent Techniques for Web Personalization and Recommender Systems, July 2008. (Cited on page 42.)

[Girvan 2002] M. Girvan and M. E. J. Newman. *Community structure in social and biological networks.* PNAS, vol. 99, no. 12, pages 7821–7826, June 2002. (Cited on page 36.)

[Godbole 2007] Namrata Godbole, Manjunath Srinivasaiah and Steven Skiena. *Large-Scale Sentiment Analysis for News and Blogs.* In Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007. (Cited on page 37.)

[Golder 2005] Scott Golder and Bernardo A. Huberman. *The Structure of Collaborative Tagging Systems.* August 2005. (Cited on pages 39 and 67.)

[Golder 2006] S Golder and B A Huberman. *Usage Pattern of Collaborative Tagging Systems.* Journal of Information Systems, vol. 32, pages 198–208, 2006. (Cited on page 48.)

[Gracia 2009] Jorge Gracia and Eduardo Mena. *Overview of a semantic disambiguation method for unstructured web contexts.* Systems Engineering, page 187, 2009. (Cited on page 48.)

[Gruber 2005] T. Gruber. *Ontology of Folksonomy: A Mash-up of Apples and Oranges.* 2005. (Cited on pages 22 and 59.)

[Guy 2009] Ido Guy, Inbal Ronen and Eric Wilcox. *Do you know?: recommending people to invite into your social network.* In Proceedings of the 13th international conference on Intelligent user interfaces, IUI '09, pages 77–86, New York, NY, USA, 2009. ACM. (Cited on pages 76 and 77.)

[Hacid 2010] Hakim Hacid, Johann Stan and Johann Daigremont. *Approche sémantique pour la préservation de la vie privée dans les médias sociaux.* pages 695–696, 2010. (Cited on page 8.)

[Hannon 2010] John Hannon, Mike Bennett and Barry Smyth. *Recommending twitter users to follow using content and collaborative filtering approaches.* In Proceedings of the fourth ACM conference on Recommender systems, RecSys '10, pages 199–206, New York, NY, USA, 2010. ACM. (Cited on page 77.)

[Herlocker 2000] Jonathan L. Herlocker, Joseph A. Konstan and John Riedl. *Explaining collaborative filtering recommendations.* In CSCW '00: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, pages 241–250, New York, NY, USA, 2000. ACM. (Cited on page 76.)

[Heymann 2006] Paul Heymann and Hector Garcia-Molina. *Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems.* Rapport technique 2006-10, Computer Science Department, April 2006. (Cited on pages 42 and 93.)

[Horowitz 2010] Damon Horowitz and Sepandar D. Kamvar. *The anatomy of a large-scale social search engine.* In Proceedings of the 19th international conference on World wide web, WWW '10, pages 431–440, New York, NY, USA, 2010. ACM. (Cited on pages 4, 9, 13 and 77.)

[Hotho 2006] Andreas Hotho, Robert Jäschke, Christoph Schmitz and Gerd Stumme. *Information Retrieval in Folksonomies: Search and Ranking.* In Proceedings of the 3rd European Semantic Web Conference, volume 4011 of *LNCS*, pages 411–426, Budva, Montenegro, June 2006. Springer. (Cited on pages 40 and 41.)

[i Cancho 2001] Ramon Ferrer i Cancho and Ricard V. Solé. *Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited.* Journal of Quantitative Linguistics, vol. 8, no. 3, pages 165–173, 2001. (Cited on page 35.)

[Java 2007] Akshay Java, Xiaodan Song, Tim Finin and Belle Tseng. *Why We Twitter: Understanding Microblogging Usage and Communities.* Procedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007, August 2007. (Cited on pages 24 and 28.)

[Jeppesen 2006] Lars Bo Jeppesen and Lars Frederiksen. *Why Do Users Contribute to Firm-Hosted User Communities? The Case of Computer-Controlled Music Instruments.* Organization Science, vol. 17, pages 45–63, January 2006. (Cited on page 17.)

[Jiang 1997] J.J. Jiang and D.W. Conrath. *Semantic similarity based on corpus statistics and lexical taxonomy.* In Proc. of the Int'l. Conf. on Research in Computational Linguistics, pages 19–33, 1997. (Cited on pages 43 and 45.)

[John 2007] Allsopp John. *hCard Contact Information Vocabulary.* 2007. [Online; accessed 20-October-2010]. (Cited on page 56.)

[Kahan 2001] José Kahan and Marja-Ritta Koivunen. *Annotea: an open RDF infrastructure for shared Web annotations.* In WWW '01: Proceedings of the 10th international conference on World Wide Web, pages 623–632, New York, NY, USA, 2001. ACM. (Cited on page 60.)

[Kajdanowicz 2010] Tomasz Kajdanowicz, Przemyslaw Kazienko and Piotr Doskocz. *A Method of Label-Dependent Feature Extraction in Social Networks.* In Jeng-Shyang Pan, Shyi-Ming Chen and Ngoc Thanh Nguyen, editeurs, ICCCI (2), volume 6422 of *Lecture Notes in Computer Science*, pages 11–21. Springer, 2010. (Cited on page 36.)

[Katifori 2007a] Akrivi Katifori, Maria Golemati, Costas Vassilakis, George Lepouras and Constantin Halatsis. *Creating an Ontology for the User Profile: Method and Applications.* In Colette Rolland, Oscar Pastor and Jean-Louis Cavarero, editeurs, RCIS, pages 407–412, 2007. (Cited on page 55.)

[Katifori 2007b] Akrivi Katifori, Maria Golemati, Costas Vassilakis, George Lepouras and Constantin Halatsis. *Creating an Ontology for the User Profile: Method and Applications.* In Colette Rolland, Oscar Pastor and Jean-Louis Cavarero, editeurs, RCIS, pages 407–412, 2007. (Cited on page 58.)

[Khelif 2008] Khaled Khelif, Fabien Gandon, Olivier Corby and Rose Dieng-Kuntz. *Using the Intension of Classes and Properties Definition in Ontologies for Word Sense Disambiguation.* In Proc. 16th International Conference on Knowledge Engineering and Knowledge Management - Knowledge Patterns, EKAW, Acitrezza, Italy, September-October 2008. (Cited on page 49.)

[Kim 2003] Hyoung R. Kim and Philip K. Chan. *Learning implicit user interest hierarchy for context in personalization.* In IUI, pages 101–108. ACM, 2003. (Cited on page 48.)

[Kim 2008a] Hak-Lae Kim, John Breslin, Sung-Kwon Yang and Hong-Gee Kim. *Social Semantic Cloud of Tag: Semantic Model for Social Tagging.* Agent and Multi-Agent Systems: Technologies and Applications, pages 83–92, 2008. (Cited on page 60.)

[Kim 2008b] Hak Lae Kim, Simon Scerri, John G. Breslin, Stefan Decker and Hong Gee Kim. *The state of the art in tag ontologies: a semantic model for tagging and folksonomies.* In DCMI '08: Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications, pages 128–137. Dublin Core Metadata Initiative, 2008. (Cited on page 54.)

[Kleinberg 1999] Jon M. Kleinberg. *Authoritative sources in a hyperlinked environment.* Journal of the ACM, vol. 46, no. 5, pages 604–632, 1999. (Cited on page 41.)

[Knerr 2006] Thomas Knerr. *Tagging Ontology - Towards a Common Ontology for Folksonomies*, 2006. http://tagont.googlecode.com/files/TagOntPaper.pdf. (Cited on page 60.)

[Lajmi 2009] Sonia Lajmi, Johann Stan, Hakim Hacid, Elod Egyed-Zsigmond and Pierre Maret. *Extended Social Tags: Identity Tags Meet Social Networks.* In IEEE, editeur, 2009 IEEE International Conference on Social Computing (SocialCom-09), August 2009. (Cited on page 159.)

[Lakshmanan 1996] Laks Lakshmanan, Fereidoon Sadri and Iyer N. Subramanian. *SchemaSQL – A Language for Interoperability in Relational Multi-database Systems.* pages 239–250. Morgan Kaufmann, 1996. (Cited on page 54.)

[Lancichinetti 2008] Andrea Lancichinetti, Santo Fortunato and Janos Kertesz. *Detecting the overlapping and hierarchical community structure of complex networks*, 2008. cite arxiv:0802.1218 Comment: 20 pages, 8 figures. Final version published on New Journal of Physics. (Cited on page 36.)

[Leacock 1998] Claudia Leacock, Martin Chodorow and George A. Miller. *Using Corpus Statistics and WordNet Relations for Sense Identification.* Computational Linguistics, vol. 24, no. 1, pages 147–165, 1998. (Cited on page 45.)

[Lee 2003] F. S. L. Lee, D. Vogel and M. Limayem. *Virtual community informatics: A review and research agenda.* Journal of Information Technology and Application, vol. 5, pages 47–61, 2003. (Cited on page 4.)

[Lee 2009] Kangpyo Lee, Hyunwoo Kim, Hyopil Shin and Hyoung-Joo Kim. *Tag Sense Disambiguation for Clarifying the Vocabulary of Social Tags.* In CSE, pages 729–734. IEEE Computer Society, 2009. (Cited on page 49.)

[Lehmann 2009] Jens Lehmann, Chris Bizer, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak and Sebastian Hellmann. *DBpedia - A Crystallization Point for the Web of Data.* Journal of Web Semantics, vol. 7, no. 3, pages 154–165, 2009. (Cited on pages 5 and 99.)

[Leroy 2010] Vincent Leroy, Berkant Barla Cambazoglu and Francesco Bonchi. *Cold start link prediction.* In Bharat Rao, Balaji Krishnapuram, Andrew Tomkins and Qiang Yang, editeurs, KDD, pages 393–402. ACM, 2010. (Cited on page 36.)

[Leskovec 2008] J. Leskovec and E. Horvitz. *Planetary-scale views on a large instant-messaging network.* 2008. (Cited on pages 35 and 36.)

[Li 2008] Xin Li, Lei Guo and Yihong Eric Zhao. *Tag-based social interest discovery.* In Jinpeng Huai, Robin Chen, Hsiao-Wuen Hon, Yunhao Liu, Wei-Ying Ma, Andrew Tomkins and Xiaodong Zhang, editeurs, WWW, pages 675–684. ACM, 2008. (Cited on page 48.)

[Li 2010] Baichuan Li and Irwin King. *Routing questions to appropriate answerers in community question answering services.* In Jimmy Huang, Nick Koudas, Gareth J. F. Jones, Xindong Wu, Kevyn Collins-Thompson and Aijun An, editeurs, CIKM, pages 1585–1588. ACM, 2010. (Cited on page 77.)

[Limpens 2009] Freddy Limpens, Alexandre Monnin, David Laniado and Fabien Gandon. *NiceTag Ontology: tags as named graphs.* 2009. (Cited on page 61.)

[Lin 2009] Ching-Yung Lin, Nan Cao, Shixia Liu, Spiros Papadimitriou, Jimeng Sun and Xifeng Yan. *SmallBlue: Social Network Analysis for Expertise Search and Collective Intelligence.* In ICDE, pages 1483–1486. IEEE, 2009. (Cited on page 77.)

[Lu 2009] Linyuan Lu and Tao Zhou. *Role of weak ties in link prediction of complex networks.* In Jun Wang, Shi Zhou and Dell Zhang, editeurs, CIKM-CNIKM, pages 55–58. ACM, 2009. (Cited on page 36.)

[Ma 2011] Yunfei Ma, Yi Zeng, Xu Ren and Ning Zhong. *User Interests Modeling Based on Multi-source Personal Information Fusion and Semantic Reasoning.* In Ning Zhong, Vic Callaghan, Ali A. Ghorbani and Bin Hu, editeurs, AMT, volume 6890 of *Lecture Notes in Computer Science*, pages 195–205. Springer, 2011. (Cited on page 101.)

[man Au Yeung 2007] Ching man Au Yeung, Nicholas Gibbins and Nigel Shadbolt. *Tag Meaning Disambiguation through Analysis of Tripartite Structure of Folksonomies.* In Web Intelligence/IAT Workshops, pages 3–6. IEEE, 2007. (Cited on page 49.)

[Maret 2004] Pierre Maret, Mark Hammond and Jacques Calmet. *Virtual Knowledge Communities for Corporate Knowledge Issues.* In Marie Pierre Gleizes, Andrea Omicini and Franco Zambonelli, editeurs, ESAW, volume 3451 of *Lecture Notes in Computer Science*, pages 33–44. Springer, 2004. (Cited on page 15.)

[Markines 2009] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho and G. Stumme. *Evaluating Similarity Measures for Emergent Semantics of Social Tagging.* In Proc. WWW, 2009. (Cited on pages 43 and 44.)

[Mathes 2004] Adam Mathes. *Folksonomies – Cooperative Classification and Communication Through Shared Metadata.* Computer Mediated Communication - LIS590CMC, December 2004. (Cited on pages 38 and 39.)

[McBride 2004] Brian McBride. *RDF Vocabulary Description Language 1.0: RDF Schema*, 2004. (Cited on page 51.)

[McGill 1983] M. McGill and G. Salton. Introduction to modern information retrieval. McGraw-Hill, 1983. (Cited on page 71.)

[McPherson 2001a] Miller McPherson, Lynn Smith-Lovin and James M Cook. *Birds of a Feather: Homophily in Social Networks.* Annual Review of Sociology, vol. 27, no. 1, pages 415–444, 2001. (Cited on page 70.)

[McPherson 2001b] Miller McPherson, Lynn Smith-Lovin and James M Cook. *Homophily in Social Networks.* Phaedrus, vol. 27, pages 415–444, 2001. (Cited on page 103.)

[Mendes 2010a] Pablo N. Mendes, Alexandre Passant and Pavan Kapanipathi. *Twarql: tapping into the wisdom of the crowd.* In Adrian Paschke, Nicola Henze and Tassilo Pellegrini, editeurs, I-SEMANTICS, ACM International Conference Proceeding Series. ACM, 2010. (Cited on page 7.)

[Mendes 2010b] Pablo N. Mendes, Alexandre Passant, Pavan Kapanipathi and Amit P. Sheth. *Linked Open Social Signals.* In Jimmy Xiangji Huang, Irwin King, Vijay V. Raghavan and Stefan Rueger, editeurs, Web Intelligence, pages 224–231. IEEE, 2010. (Cited on page 24.)

[Michelson 2010] Matthew Michelson and Sofus A. Macskassy. *Discovering users' topics of interest on twitter: a first look.* In Roberto Basili, Daniel P. Lopresti, Christoph Ringlstetter, Shourya Roy, Klaus U. Schulz and L. Venkata Subramaniam, editeurs, AND, pages 73–80. ACM, 2010. (Cited on page 71.)

[Mihalcea 2006] Rada Mihalcea, Courtney Corley and Carlo Strapparava. *Corpus-based and knowledge-based measures of text semantic similarity.* In In AAAI'06, pages 775–780, 2006. (Cited on page 45.)

[Mika 2005] Peter Mika. *Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks.* Journal of Web Semantics, vol. 3, pages 211–223, 2005. (Cited on page 75.)

[Mika 2007a] Peter Mika. *Ontologies are us: A unified model of social networks and semantics.* Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5, no. 1, pages 5 – 15, 2007. Selected Papers from the International Semantic Web Conference, International Semantic Web Conference (ISWC2005). (Cited on pages 37, 40, 41, 42, 46 and 59.)

[Mika 2007b] Peter Mika. Social networks and the semantic web, volume 5 of *Semantic Web And Beyond Computing for Human Experience.* Springer, 2007. (Cited on page 51.)

[Miles 2008] Alistair Miles and Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference.* August 2008. (Cited on page 61.)

[Milicic 2008] Vuk Milicic. *Case Study: Semantic tags.* Website, December 2008. `http://www.w3.org/2001/sw/sweo/public/UseCases/Faviki/Faviki.pdf`. (Cited on page 38.)

[Miller 1995] G. A. Miller. *WordNet: a lexical database for English.* Communications of the ACM, vol. 38, no. 11, pages 39–41, 1995. (Cited on page 48.)

[Morris 2010] Meredith Ringel Morris, Jaime Teevan and Katrina Panovich. *What do people ask their social networks, and why?: a survey study of status message q&a behavior.* In Proceedings of the 28th international conference on Human factors in computing systems, CHI '10, pages 1739–1748, New York, NY, USA, 2010. ACM. (Cited on page 4.)

[Naaman 2010a] Mor Naaman, Jeffrey Boase and Chih H. Lai. *Is it really about me?: message content in social awareness streams.* pages 189–192, 2010. (Cited on pages 22, 28 and 93.)

[Naaman 2010b] Mor Naaman, Jeffrey Boase and Chih-Hui Lai. *Is it really about me?: message content in social awareness streams.* In Proceedings of the 2010 ACM conference on Computer supported cooperative work, CSCW '10, pages 189–192, New York, NY, USA, 2010. ACM. (Cited on page 24.)

[Nadeau 2007] David Nadeau and Satoshi Sekine. *A survey of named entity recognition and classification.* Linguisticae Investigationes, vol. 30, pages 3–26, 2007. (Cited on page 47.)

[Nambisan 2002] Satish Nambisan. *DESIGNING VIRTUAL CUSTOMER ENVIRONMENTS FOR NEW PRODUCT DEVELOPMENT: TOWARD A THEORY.* Academy of Management Review, vol. 27, no. 3, pages 392–413+, 2002. (Cited on page 17.)

[Newman 2005] Richard Newman, Danny Ayers and Seth Russell. *Tag Ontology*, December 2005. http://www.holygoat.co.uk/owl/redwood/0.1/tags/. (Cited on page 59.)

[Newman 2006] Mark Newman, Albert László Barabási and Duncan J. Watts. The structure and dynamics of networks. Princeton University Press, 2006. (Cited on pages 35 and 36.)

[Noor 2009] Salma Noor and Kirk Martinez. *Using Social Data as Context for Making Recommendations: An Ontology based Approach.* Proceedings of the 1st Workshop on Context Information and OntologiesCIAO 2009 Jun 1 Herakleion Greece, page 7, 2009. (Cited on page 49.)

[Page 1999] Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd. *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120. (Cited on page 9.)

[Palla 2005] Gergely Palla, Imre Derenyi, Illes Farkas and Tamas Vicsek. *Uncovering the overlapping community structure of complex networks in nature and society.* Nature, vol. 435, no. 7043, pages 814–818, jun 2005. (Cited on page 36.)

[Passant 2008a] A. Passant and P. Laublet. *Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data.* Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, Apr, 2008. (Cited on pages 48 and 60.)

[Passant 2008b] Alexandre Passant. *LODr - A Linking Open Data Tagging System.* In Proceedings of the First Social Data on the Web Workshop (SDoW2008), Karlsruhe, Germany, October 27 2008. (Cited on page 62.)

[Passant 2008c] Alexandre Passant and Philippe Laublet. *Combining Structure and Semantics for Ontology-Based Corporate Wikis.* In Business Information Systems, 11th International Conference, BIS 2008, Innsbruck, Austria, May 2008, Lecture Notes in Business Information Processing, pages 58–69. Springer-Verlag, 2008. (Cited on page 38.)

[Plangprasopchok 2009] Anon Plangprasopchok and Kristina Lerman. *Constructing folksonomies from user-specified relations on flickr.* In Juan Quemada, Gonzalo León, Yoëlle S. Maarek and Wolfgang Nejdl, editeurs, WWW, pages 781–790. ACM, 2009. (Cited on page 43.)

[Porter 1980] M. F. Porter. *An algorithm for suffix stripping.* Program, vol. 14, no. 3, pages 130–137, July 1980. (Cited on page 47.)

[Preece 2004] Jennifer J. Preece, Blair Nonnecke and Dorine Andrews. *The top five reasons for lurking: improving community experiences for everyone.* Computers in Human Behavior, vol. 20, no. 2, pages 201–223, 2004. (Cited on page 17.)

[Razavi 2009] Maryam Najafian Razavi and Lee Iverson. *Improving personal privacy in social systems with people-tagging.* In GROUP, pages 11–20, 2009. (Cited on page 74.)

[Resnik 1999] P. Resnik. *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language.* Journal of Artificial Intelligence Research, vol. 11, no. 1, pages 95–130, 1999. (Cited on page 45.)

[Rheingold 2000] Howard Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier, Revised Edition.* 2000. (Cited on page 17.)

[Robinson 2011] Julien Robinson, Johann Stan and Myriam Ribiere. *Using Linked Data to Reduce Learning Latency for e-Book Readers.* vol. 717, 2011. (Cited on page 8.)

[Rousseau 2004] Boris Rousseau, Parisch Browne, Paul Malone, Paul Foster and Venura Mendis. *Personalised Resource Discovery Searching over Multiple Repository Types: Using User and Information Provider Profiling.* In ICEIS (5), pages 35–43, 2004. (Cited on page 57.)

[Saint-Onge 2003] Hubert Saint-Onge and Debra Wallace. Leveraging communities of practice for strategic advantage. Butterworth-Heinemann, Amsterdam [u.a.], 2003. Hubert Saint-Onge; Debra Wallace. ill 24 cm. (Cited on page 17.)

[Sakaki 2010] Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo. *Earthquake shakes Twitter users: real-time event detection by social sensors.* In Proceedings of the 19th international conference on World wide web, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM. (Cited on page 24.)

[Salton 1975] Gerard Salton, Anita Wong and Chung-Shu Yang. *A Vector Space Model for Automatic Indexing.* Communications of the ACM, vol. 18, no. 11, pages 613–620, 1975. The paper where vector space model for IR was introduced. (Cited on page 125.)

[Schmitz 2006] Patrick Schmitz. *Inducing Ontology from Flickr Tags.* In Proceedings of the Workshop on Collaborative Tagging at WWW2006, Edinburgh, Scotland, May 2006. (Cited on page 42.)

[Shadbolt 2006] Nigel Shadbolt, Tim Berners-Lee and Wendy Hall. *The Semantic Web Revisited.* IEEE Intelligent Systems, vol. 21, pages 96–101, May 2006. (Cited on page 54.)

[Shahabi 1997] Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi and Vishal Shah. *Knowledge Discovery from Users Web-Page Navigation.* In RIDE, pages 0–, 1997. (Cited on page 71.)

[Shannon 1948] Claude E. Shannon. *A mathematical theory of Communication.* The Bell system technical journal, vol. 27, pages 379–423, July 1948. (Cited on page 123.)

[Stan 2009] Johann Stan, Elod Egyed-Zsigmond, Pierre Maret and Johann Daigremont. *Efficient Reachability Management With A Semantic User Profile Framework.* In Personalization in Mobile and Pervasive Computing (workshop at User Modeling, Adaptation, and Personalization conference - UMAP), June 2009. (Cited on page 165.)

[Stan 2011a] Stan, Do and Maret. *User Interaction Profiles for Better People Recommendation.* International Conference on Advances in Social Network Analysis and Mining, vol. 0, pages 206–211, 2011. (Cited on page 7.)

[Stan 2011b] J. Stan, Ribiere M., Picault J., Natarianni L. and Marie N. *Semantic-Awareness for a Useful Digital Life.* Social Network Mining, Analysis and Research Trends: Techniques and Applications, 2011. (Cited on page 9.)

[Stankovic 2008] M. Stankovic. *Modeling Online Presence.* In J. Breslin, U. Bojars, A. Passant and S. Fernandez, editeurs, Proceedings of the First Social Data on the Web Workshop, Karlsruhe, Germany, volume 405, Karlsruhe, Germany, 10 2008. CEUR Workshop Proceedings. (Cited on page 58.)

[Suchanek 2008] Fabian M. Suchanek, Milan Vojnovic and Dinan Gunawardena. *Social tags: meaning and suggestions.* In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM '08, pages 223–232, New York, NY, USA, 2008. ACM. (Cited on page 38.)

[Sugiyama 2004] Kazunari Sugiyama, Kenji Hatano and Masatoshi Yoshikawa. *Adaptive web search based on user profile constructed without any effort from users.* In WWW '04: Proceedings of the 13th international conference on World Wide Web, pages 675–684, New York, NY, USA, 2004. ACM Press. (Cited on page 71.)

[Sun 2011] Ming Sun and Jerome R. Bellegarda. *Improved pos tagging for text-to-speech synthesis.* In ICASSP, pages 5384–5387. IEEE, 2011. (Cited on page 122.)

[Sutterer 2008] Michael Sutterer, Olaf Droegehorn and Klaus David. *UPOS: User Profile Ontology with Situation-Dependent Preferences Support.* In ACHI '08: Proceedings of the First International Conference on Advances in Computer-Human Interaction, pages 230–235, Washington, DC, USA, 2008. IEEE Computer Society. (Cited on page 58.)

[Szomszor 2008] Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara and Nigel Shadbolt. *Semantic modelling of user interests based on cross-folksonomy analysis.* Architecture, vol. 5318, pages 632–648, 2008. (Cited on pages 48 and 80.)

[Teevan 2005] Jaime Teevan, Susan T. Dumais and Eric Horvitz. *Personalizing search via automated analysis of interests and activities.* In SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pages 449–456, New York, NY, USA, 2005. ACM. (Cited on page 55.)

[Terveen 2005] Loren Terveen and David W. McDonald. *Social matching: A framework and research agenda.* ACM Trans. Comput.-Hum. Interact., vol. 12, no. 3, pages 401–434, 2005. (Cited on page 151.)

[The W3C Consortion 2010] The W3C Consortion. *Web Ontology Language (OWL)*, 2010. (Cited on page 51.)

[Trajkova 2004] Joana Trajkova and Susan Gauch. *Improving Ontology-Based User Profiles.* In Christian Fluhr, Gregory Grefenstette and W. Bruce Croft, editeurs, RIAO, pages 380–390. CID, 2004. (Cited on page 55.)

[Tseng 2007] Belle L. Tseng. *Blog analysis and mining technologies to summarize the wisdom of crowds.* In Proceedings of the 8th international workshop on Multimedia data mining: (associated with the ACM SIGKDD 2007), MDM '07, pages 3:1–3:1, New York, NY, USA, 2007. ACM. (Cited on page 37.)

[Vallet 2007] David Vallet, Pablo Castells, Miriam Fernández, Phivos Mylonas and Yannis S. Avrithis. *Personalized Content Retrieval in Context Using Ontological Knowledge.* IEEE Trans. Circuits Syst. Video Techn., vol. 17, no. 3, pages 336–346, 2007. (Cited on pages 69 and 71.)

[Vildjiounaite 2007] Elena Vildjiounaite, Otilia Kocsis, Vesa Kyllönen and Basilis Kladis. *Context-Dependent User Modelling for Smart Homes.* In Cristina Conati, Kathleen F. McCoy and Georgios Paliouras, editeurs, User Modeling, volume 4511 of *Lecture Notes in Computer Science*, pages 345–349. Springer, 2007. (Cited on page 58.)

[von Hessling 2004] A. von Hessling, T. Kleemann and A. Sinner. *Semantic User Profiles and their Applications in a Mobile Environment.* In Artificial Intelligence in Mobile Systems 2004, 2004. (Cited on page 57.)

[Wagner 2010] C. Wagner and M. Strohmaier. *The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams.* In Proc. of the Semantic Search 2010 Workshop (SemSearch2010), april 2010. (Cited on pages 7, 46 and 71.)

[Wang 2003] Yang Wang, Deepayan Chakrabarti, Chenxi Wang and Christos Faloutsos. *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint.* In SRDS, pages 25–34. IEEE Computer Society, 2003. (Cited on page 36.)

[Wang 2007] Fei-Yue Wang, Kathleen M. Carley, Daniel Zeng and Wenji Mao. *Social Computing: From Social Informatics to Social Intelligence.* IEEE Intelligent Systems, vol. 22, no. 2, pages 79–83, 2007. (Cited on page 7.)

[Wellman 1996] Barry Wellman. *For a Social Network Analysis of Computer Networks: A Sociological Perspective on Collaborative work and Virtual Community.* no. 1-11, 1996. (Cited on page 16.)

[Williams 2000] R. L. Williams and J. Cothrel. *Four smart ways to run online communities.* Sloan Management Review, vol. 41, no. 4, pages 81–92, 2000. (Cited on page 17.)

[Windley 2005] Phillip J. Windley. Digital identity. O'Reilly, 2005. (Cited on page 55.)

[Wu 1994] Zhibiao Wu and Martha Palmer. *Verb Semantics And Lexical Selection.* In Proc. of the 32nd annual meeting on Association for Computational Linguistics, pages 133–138, 1994. (Cited on page 45.)

[Wu 2006] H. Wu, M. Zubair and K. Maly. *Harvesting social knowledge from folksonomies.* In Proceedings of the seventeenth conference on Hypertext and hypermedia, pages 111–114. ACM Press New York, NY, USA, 2006. (Cited on page 40.)

[Xi 2004] Wensi Xi, Benyu Zhang, Zheng Chen, Yizhou Lu, Shuicheng Yan, Wei-Ying Ma and Edward A. Fox. *Link fusion: a unified link analysis framework for multi-type interrelated data objects.* In WWW '04: Proceedings of the 13th international conference on World Wide Web, pages 319–327, New York, NY, USA, 2004. ACM Press. (Cited on page 42.)

[Xu 2005] Jennifer Jie Xu and Hsinchun Chen. *Criminal network analysis and visualization.* Commun. ACM, vol. 48, no. 6, pages 100–107, 2005. (Cited on page 36.)

[Yeh 2003] Iwei Yeh, Peter D. Karp, Natalya Fridman Noy and Russ B. Altman. *Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO).* Bioinformatics, vol. 19, no. 2, pages 241–248, 2003. (Cited on page 22.)

[Zhou 2007] Mianwei Zhou, Shenghua Bao, Xian Wu and Yong Yu. *An Unsupervised Model for Exploring Hierarchical Semantics from Social Annotations.* In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber and Philippe Cudré-Mauroux, editeurs, ISWC/ASWC, volume 4825 of *Lecture Notes in Computer Science*, pages 680–693. Springer, 2007. (Cited on pages 42, 43 and 80.)