Gestion des connaissances et communication médiatisée : traçabilité et structuration des messages professionnels

François Rauscher

Thèse
de doctorat
de l'UTT

François RAUSCHER

Gestion des connaissances et communication médiatisée : traçabilité et structuration des messages professionnels

Spécialité :
Ingénierie Sociotechnique des Connaissances, des Réseaux et du Développement Durable

2016TROY0032

Année 2016

utt
université de technologie
Troyes

THESE

pour l'obtention du grade de

DOCTEUR de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

Spécialité : INGENIERIE SOCIOTECHNIQUE DES CONNAISSANCES, DES RESEAUX ET DU DEVELOPPEMENT DURABLE

présentée et soutenue par

François RAUSCHER

le 13 octobre 2016

Gestion des connaissances et communication médiatisée : traçabilité et structuration des messages professionnels

JURY

M. E. CAILLAUD	PROFESSEUR DES UNIVERSITES	Président
M. H. ATIFI	MAITRE DE CONFERENCES	Directeur de thèse
M. É. BONJOUR	PROFESSEUR DES UNIVERSITES	Examinateur
Mme V. GOEPP	MAITRE DE CONFERENCES - HDR	Rapporteur
Mme M. LEWKOWICZ	PROFESSEUR DES UNIVERSITES	Examinateur
Mme N. MATTA	PROFESSEURE UTT	Directrice de thèse
M. D. MONTICOLO	MAITRE DE CONFERENCES - HDR	Rapporteur

RESUME / ABSTRACT

Même si le capital immatériel représente une part de plus en plus importante de la valeur de nos organisations, il n'est pas toujours possible de stocker, tracer ou capturer les connaissances et les expertises, par exemple dans des projets de taille moyenne. Le courrier électronique est encore largement utilisé dans les projets d'entreprise en particulier entre les équipes géographiquement dispersées. Dans cette étude, nous présentons une nouvelle approche pour détecter les zones à l'intérieur de courriels professionnels où des éléments de connaissances sont susceptibles de se trouver. Nous définissons un contexte étendu en tenant compte non seulement du contenu du courrier électronique et de ses métadonnées, mais également des compétences et des rôles des utilisateurs. Egalement l'analyse pragmatique linguistique est mêlée aux techniques usuelles du traitement de langage naturel. Après avoir décrit notre méthode KTR et notre modèle, nous l'appliquons à un corpus réel d'entreprise et évaluons les résultats en fonction des algorithmes d'apprentissage, de filtrage et de recherche.

Even if intangible capital represents an increasingly important part of the value of our enterprises, it's not always possible to store, trace or capture knowledge and expertise, for instance in middle sized projects. Email it still widely used in professional projects especially among geographically distributed teams. In this study we present a novel approach to detect zones inside business emails where elements of knowledge are likely to be found. We define an enhanced context taking into account not only the email content and metadata but also the competencies of the users and their roles. Also linguistic pragmatic analysis is added to usual natural language processing techniques. After describing our model and method KTR, we apply it to a real life corpus and evaluate the results based on machine learning, filtering and information retrieval algorithms.

Mots clés / Key Words : Gestion des connaissances, Messageries électroniques, Traçabilité, Communication dans les organisations, Analyse de contenu (communication), Logiciels, Pragmatique, Recherche de l'information /Knowledge management, Electronic mail systems, Traceability, Communication in organizations, content analysis (communication), software, Pragmatics, Information Retrieval

REMERCIEMENTS

Je tiens d'abord à remercier Virgine Goepp et Davy Monticolo d'avoir accepté de rapporter ma thèse, et de l'intérêt qu'ils ont porté à mon travail. C'est un grand honneur, et je les remercie très sincèrement pour les retours de grande qualité qu'ils m'ont donnés à propos du manuscrit. Je remercie également Eric Bonjour, Emmanuel Caillaud et Myriam Lewkowicz d'avoir accepté de faire partie du jury et d'assister à la présentation de ce travail.

Je remercie également Nada Matta d'avoir initié et codirigé cette thèse. Toujours présente, elle a su me guider dans mes travaux, m'apprendre l'exigence de la démarche scientifique rigoureuse, me faire profiter de son expérience et de ses connaissances tout en me laissant explorer par moi-même.

J'admire ses qualités humaines, sa disponibilité, sa bonne humeur et surtout sa grande bonté. Les nombreux échanges que nous avons eu au cours de cette thèse et pendant les conférences internationales m'ont enrichi et ont modifié mon regard sur le monde et pas seulement scientifique. Je lui en suis profondément reconnaissant.

Je remercie aussi sincèrement Hassan Atifi d'avoir codirigé cette recherche. Il a eu la patience de me voir évoluer dans la découverte de la linguistique et de la pragmatique de la communication. Il m'a fait profiter de ses compétences expertes et a toujours été présent pour des conseils et relectures. Ses remarques et parfois remontrances dans mes premières recherches m'ont permis de ne pas m'égarer, de conserver une pensée critique et là encore de renforcer ma démarche scientifique.

Ces quatre ans de travail à leurs côtés ont été très intenses et éminemment formateurs.

Je tiens par ailleurs à exprimer toute ma reconnaissance à Jean-Baptiste Jouannaud du Groupe Infopro Digital à l'origine de notre corpus pour le temps et la confiance qu'il nous a accordé.

Merci aussi à toute l'équipe de Tech Cico, au sein de laquelle j'ai fait ma thèse. Je remercie également Stéphanie et Christine pour leur patience, leur disponibilité et leur aide efficace pour mes démarches administratives. De manière générale je souhaite

remercier tout le personnel de l'Ecole doctorale qui a fait de ce lieu un endroit où il est agréable de venir pour travailler.

Je remercie Jason et Amina mes compagnons de thèse et de conférences avec qui j'ai eu de nombreuses discussions et pour les visites et les repas bien agréables.

Au hasard des conférences je tiens aussi à remercier chaleureusement des personnes avec qui j'ai eu des discussions pour le rôle moteur qu'ils ont eu, peut-être sans le savoir, dans mes recherches : Gloria Origgi, Eunika Mercier-Laurent, Fabien Gandon, Christophe Lejeune et tous les intervenants de la classe d'été Ermites 2014.

Merci à mes amis Jean-Michel, les Michels, Daniel, Guy et Elzéar pour leur intelligence et leur érudition lors de nos nombreux échanges qui ont participé à ma réflexion.

Merci bien entendu à Christian Mariusse, fidèle parmi les fidèles, qui m'a fait confiance à mes débuts et à soutenu mes recherches même de loin et pendant de nombreuses années. Son amitié et nos longues soirées de discussion me seront toujours précieuses.

Je remercie aussi bien sur tous ceux dont la présence et le soutien au quotidien ont été primordiaux pour moi. Mes proches : Florence, Michel, Jean, Victor, et mon ami Adem pour son bon sens et son énergie. Enfin mes parents Mireille et André qui m'ont appris l'étrange alliance possible entre les mathématiques et la littérature et qui ont cru en moi toutes ces années.

Et bien sur un grand merci à Isabelle et ma petite Alix.

TABLE DES MATIERES

LISTE DES FIGURES

INTRODUCTION

Contexte général

La gestion des connaissances a connu un essor formidable au sein des organisations dans les années 2000 après les travaux et découvertes de la décennie précédente (Steels, 1993), (Van Heijst, 1996). En effet, avec l'avènement d'internet dans les années 1990, la diffusion et le partage d'informations se sont fortement accrus, ainsi que l'ambition de pouvoir appliquer cela aux connaissances. Les entreprises ont réellement pris conscience de la valeur stratégique des connaissances en tant que capital immatériel (Grundstein 1995).

De nos jours, bien que l'effet de mode du « Knowledge Management » se soit un peu amoindri, la gestion des connaissances est devenue cruciale pour les entreprises. Les industriels ont compris les avantages concurrentiels que l'on pouvait retirer de la capitalisation des connaissances. Celle-ci peut prendre diverses formes selon la taille de l'organisation. Pour les TPE et PME, elle est embryonnaire et se résume souvent à de l'archivage électronique ou un simple portail collaboratif (comme SharePoint). Pour les entreprises de plus grande taille et les groupes internationaux, des programmes plus conséquents ont été mis en place pour gérer le capital immatériel.

Des méthodes de recueil, de modélisation et de stockage ont été développées dans les laboratoires (CommonKAD, MASK, KOD par exemple.) et mises en application dans

des grands groupes (Areva, PSA, Renault, l'Aérospatiale, etc..). Essentiellement basée sur le recueil de savoir d'experts et avec une grande partie d'opérations manuelles et d'interviews par des cogniticiens, ces initiatives ont permis la mise en place de livres de connaissance et/ou de logiciel de partage, de localisation d'experts. Le but recherché est toujours le même : Au cours d'une des opérations de l'entreprise dans le cadre de son fonctionnement du savoir-faire a été créé et/ou utilisé. Cette connaissance ne doit pas être perdue, car elle peut servir dans des domaines similaires pour l'entreprise, et parfois dans des domaines différents (par analogie ou conceptualisation), ou encore en termes de traçabilité pour déterminer par qui et pourquoi une décision a été prise, et quel mode opératoire en a découlé.

Cette approche est rassurante pour les grands groupes mais a parfois rencontré ses limites car cette chaine de valorisation des connaissances comme l'indique Nonaka (Nonaka, 1991) a un cycle d'existence, il faut la faire vivre et gérer son obsolescence programmée. En outre, si la taille de l'organisation est conséquente, soit on a affaire à une mémoire trop générique et délicate à exploiter, soit à une fédération de différentes mémoires ayant traits à des activités mais sans réel lien entre elles (à part certains acteurs). Souvent des ontologies de domaines sont associées à ces mémoires et en étant plus restreintes permettent une exploitation plus efficace mais au détriment de la généralisation.

La mémoire de projet est apparue dans ce contexte, plus centrée, plus spécifique qu'une mémoire d'entreprise, elle avait pour but évident la traçabilité et la réutilisation en cas de projet similaire ou de suite. La mémoire de projet (Matta, 1999) utilise comme matériaux les données du projet, les produits du projet, les intervenants (leur rôle, compétences dans l'organisation, les échanges et réunions, communications électronique, téléphonique) les documents. On se heurte parfois à un effet d'échelle inverse à celui des mémoires d'entreprise : il est parfois difficile de mettre en place un procédé, un logiciel et des interviews de recueil pour un projet réunissant moins d'une dizaine de personnes. Des outils ont été développés pour annoter les réunions (Matta, 2010) mais la taille des équipes rend couteux et difficile le recueil systématique (surtout a posteriori).

Enjeux scientifiques

Dans le cadre de cette étude, nous sommes intéressé aux entreprises de taille moyenne (>1000 personnes) ayant des projets impliquant des équipes géographiquement distribuées, ou pratiquant le télétravail (de plus en plus fréquent, on note selon une étude américaine[1] une croissance de 102.1% entre 2005 et 2014, il est utilisé par exemple sur les projets de développement informatique (12% dans le domaine des TIC selon une étude Obergo 2015[2]).

Ces entreprises nous ayant consulté avec des besoins identifiés en termes de traçabilité et de réutilisation de connaissances mais en ayant des ressources restreintes. Les projets en question étaient terminés, il restait les livrables/les produits, les données de la gestion de projet (planning, documentation, spécifications), la liste des participants et leur compétences, et l'ensemble de leur échange électroniques avec les documents afférents. A noter qu'il n'y avait pas d'enregistrement (audio ou vidéo) des réunions, ni des conversations téléphoniques/ vidéoconférences.

La question était donc de déterminer si, dans un corpus d'email professionnels, avec les données de contexte du projet (intervenants, organisation, document, produit résultat), nous pouvions localiser des « connaissances » dans ces emails, si oui de quel type, et ce procédé était-il automatisable et surtout comment les réutiliser a posteriori ?

L'email est un moyen de communication qui est apparu en même temps que le World Wide Web. Il n'a guère évolué depuis mais malgré sa mort annoncée à maintes reprises (au profit des réseaux sociaux ou des messageries instantanée), il est toujours présent et primordial dans la vie des entreprises. Quelques études se sont intéressées à l'email d'un point de vue purement linguistique (Barron, 1998) ou encore communicationnel avec la CMDA (Computer-Mediated Discourse Analysis) (Herring, 2004). Principalement les travaux sur l'email se sont focalisés sur la gestion du spam, la classification, ou les topics (cluster). Cependant les travaux de S Herring

[1] http://globalworkplaceanalytics.com/

[2] http://www.ergostressie.com/

ont permis de mettre en valeur des caractéristiques de l'email qui bien qu'asynchrone et distant pouvait être assimilé à une forme de discours et donc rendre possible l'application de techniques d'analyse conversationnelle.

En effet la plupart du temps les techniques employées autour de l'email était issues du TAL (Traitement Automatique du Langage en anglais NLP (Natural Language Processing)) même si la taille des messages s'adaptait parfois mal aux contraintes de l'analyse statistique. On notera les travaux de Bickel (2005) et l'étude de Guzella (2009). Cependant certains auteurs se sont tournés vers une autre approche via une branche de linguistique : la pragmatique.

La pragmatique linguistique, issue des travaux de Searle (1969) et Austin (1962), privilégie non pas une approche d'analyse grammaticale, mais se centre sur les actes de langage (speech acts), sur l'aspect communication et le contexte. Un acte de langage est un acte que l'on effectue en le disant, on catégorise ces actes suivant leur motivation (ordonner, promettre, demander, etc.) Par rapport à une simple analyse syntaxique (NLP) ou même sémantique, prendre en compte le but recherché par le locuteur lors d'un énoncé facilite son interprétation. Notre étude est plus centrée sur la requête (directe ou indirecte) car elle est présente lors de l'énoncé d'une problématique dans un projet.

Tout d'abord nous avons choisi de nous focaliser sur la résolution de problème, car c'est là où des connaissances utiles pour l'entreprise sont le plus susceptibles d'être employées ou créées. Et plus particulièrement sur les problèmes de type indéfinis ou « wicked problems » (Buckingham-Shum, 1997), (Conklin et Weil, 1997) où la connaissance collaborative se manifeste naturellement.

Nous avons développé une grille d'analyse qui inclue les données d'un projet (organisation, intervenant et leur rôle et compétences, échanges d'email, produit et document associés) qui nous a permis de mettre en valeur des échanges de type problem-solving, et les motifs (patterns) associés (type d'intervenant, type d'acte de langage, contexte projet). Ces « patterns » bien que présentant des proximités avec des règles de type si/alors se révèlent de nature non linéaires (étant donné le nombre de paramètres). Il paraissait naturel de s'intéresser à une approche utilisant le Machine Learning (Apprentissage automatique).

Après validation de cette grille sur un projet type, nous avons alors évalué des algorithmes de Machine Learning afin d'appliquer cette grille sur des corpus de grande taille. Cette approche permet de repérer des « traces des connaissance » (analogues à l'email zoning de Lampert (2009)). La solution retenue est un système hybride empruntant la fois à l'ingénierie des connaissances, à la pragmatique, et au Machine Learning.

Enfin une fois les traces de connaissance repérées via un algorithme, il restait la partie finale, comment restituer ces connaissances aux utilisateurs et dans quel contexte. Problématique proche de l'IR (information retrieval) mais contextuelle.

Notre solution complète KTR (Knowledge Traces Retrieval) privilégie un mode actif de l'utilisateur, il va présenter une requête, notre système va enrichir le contexte (via les compétences et les topics), va transmettre cette requête augmentée à notre système de zonage de connaissance (via la pragmatique et le Machine Learning) pour restituer les traces de connaissance et les documents associés à l'utilisateur.

Contexte Industriel

Cette recherche s'est effectuée en partenariat avec le groupe INFOPRO DIGITAL (groupe de presse français) éditeur de nombreux magazines et ouvrages. Pour le groupe INFOPRO la gestion des connaissances et la traçabilité sur un projet est primordiale car vu le nombre de collaborateurs et les mouvements de personnels (interne ou externe), il est relativement fréquent d'avoir des questions sans réponses sur un projet en production depuis 3 à 5 ans. Des corpus réels ont été fournis pour nos études.

Organisation du mémoire

Dans un premier chapitre, nous ferons quelques rappels sur l'ingénierie des connaissances, la gestion des projets, les moyens de recueil, etc... Et en particulier la mémoire de projet.

Le second chapitre sera plus particulièrement centré sur le type de connaissances associé aux problèmes non définis, et aux systèmes permettant aux utilisateurs de base de cas, et sur la traçabilité. Au chapitre 3 nous nous intéresserons à la communication médiatisée, puis à l'email avec ses particularités. Notamment nous verrons dans quelle mesure l'email est assimilable à un discours et comment les techniques d'analyse conversationnelle peuvent s'appliquer. Nous verrons que les caractéristiques de l'email ne favorisent pas une analyse purement textuelle et statistique et que des approches différentes ont été tentées, par exemple la pragmatique linguistique, que nous présenterons plus en détail. Au chapitre 4, nous nous intéresserons aux approches de traitement du langage naturel, en particulier les applications autour de l'email, et aux manières d'associer les emails à des topics (thèmes).

Puis nous présenterons notre approche au chapitre 5, une méthode de découverte de traces de connaissance dans les emails basée sur le machine Learning, puis une formalisation de la restitution de connaissance.

Au chapitre 6, nous appliquerons cette méthode à un corpus issu d'un véritable projet industriel de notre partenaire INFOPRO. Ce projet s'étalait sur 2 ans et englobait une vingtaine de personnes. Nous étudierons les performances de notre approche hybride avec des algorithmes de Machine Learning (Support Vector Machine), puis les performances globales de notre système sur ce corpus.

Enfin nous conclurons sur les perspectives en termes de conservation, de transfert de connaissance et de traçabilité d'un tel système. Les extensions envisageables concernant les algorithmes supports et l'application à des corpus de nature différentes (système de messages) et d'autres modes de restitution pour l'utilisateur.

Chapitre 1. GESTION DES CONNAISSANCES ET GESTION DE PROJET

La gestion des connaissances (Knowledge Management ou KM) a pris une place croissante depuis la dernière décennie dans la vie des entreprises. Loin d'être un simple effet de mode, cet engouement est l'aboutissement d'un long cheminement dans la manière dont l'homme et les organisations traitent les données et informations essentielles à leur survie. Dans sa définition classique (Platon, le Théétète, 1970), la connaissance est une « croyance vraie et justifiée ». Au début sa transmission s'effectuait principalement via un savoir livresque. On notera également les corporations, par exemple les maçons ou les charpentiers, qui se transmettaient des savoir-faire et des secrets de métier (Coornaert, 1966). Cependant ce n'est qu'avec la Révolution industrielle et la taylorisation du travail que les entreprises vont prendre conscience de la nécessité de consigner les savoirs de fabrication et de gérer la connaissance de production. Les crises successives et la demande de productivité accrue vont mettre en évidence ce besoin. En effet dans un monde en perpétuel changement avec une demande d'innovation constante, il est devenu impossible de se reposer sur des savoirs acquis, des recettes existantes sans risquer de perdre sa position et de péricliter. Aussi les organisations sont obligées de mettre en place des systèmes

afin de déterminer quelles sont leur connaissances, comment et lesquelles conserver, quelles personnes les détiennent, comment en acquérir de nouvelles, et enfin comment les diffuser.

Ce type de système et de politique n'est plus un luxe réservé à de grands groupes, c'est une nécessité impérieuse pour les industriels afin d'assurer leur survie et définir leur stratégie. Si c'est un fait établi à l'échelle d'une entreprise, c'est parfois plus délicat à mettre en œuvre dans un cadre plus réduit, par exemple celui d'un simple projet ne regroupant qu'un nombre moindre d'employés.

Notre étude concerne les emails professionnels d'une entreprise échangés lors d'un projet et les connaissances utiles qu'ils pourraient contenir. Nous allons donc présenter rapidement la capitalisation des connaissances en entreprise, puis nous nous intéresserons au Projet, sa gestion, ses composants et ses acteurs. Enfin nous examinerons plus précisément la mémoire de Projet.

1.1 La Gestion des connaissances (KM) en entreprise

Tisseyre (2000) définit le Knowledge Management comme « la gestion consciente, coordonnée et opérationnelle de l'ensemble des informations, connaissances et savoir faire des membres d'une organisation au service de celle-ci ». Pour Grundstein (2002) et dans le cadre de l'entreprise la capitalisation des connaissances répond à deux finalités organisationnelles complémentaires : une finalité patrimoniale et une finalité d'innovation durable. Ces finalités répondent aux 5 facettes de la problématique illustrées dans la figure 1.1.

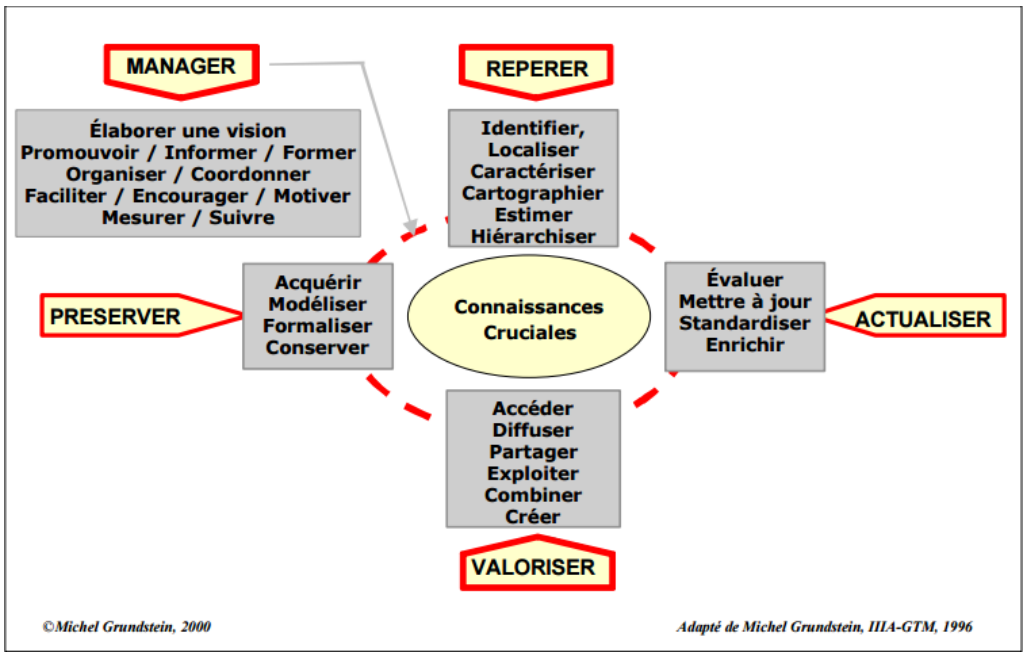


Figure 1.1 : Problématique de la capitalisation des connaissances (Grundstein, 2002)

La finalité patrimoniale s'explique par les besoins de localisation, de diffusion, d'actualisation et de préservation de la connaissance (par exemple causé par la mobilité interne ou les départs à la retraite de la génération des baby-boomers). L'innovation durable est devenue une nécessité face à la mondialisation de l'économie et à une concurrence accrue, elle concerne la création de connaissance au niveau individuel et son intégration au niveau organisationnel. Ces constatations ont fait sortir le KM des centres de recherches et développement, et c'est l'ensemble des départements de l'entreprise (marketing, production, ressources humaine, juridique, etc..) qui sont concernés. La connaissance devient une ressource à gérer au même titre que toutes les ressources matérielles d'où la mise en place de systèmes de gestion des connaissances (Dieng-Kuntz, 2002).

Dans les années 80, les travaux sur l'intelligence artificielle, les systèmes experts et l'essor de l'informatique ont permis un meilleur traitement des données et de l'information. On s'applique à essayer de stocker et codifier les connaissances issues de documents, puis d'experts. Cela donnera l'ingénierie des connaissances (Aussenac, 1994) qui a connu son apogée dans les années 90 avec la réalisation de

systèmes de connaissances puis de bases de connaissances et la pratique des retours d'expérience (Malvache and Prieur, 1993).

Mais cette approche ne concerne que l'information à valeur ajoutée structurée (sous forme de documents), soit stockée (systèmes experts servant à capitaliser et mettre à disposition des savoirs individuels devenus rares ou difficilement accessibles), soit diffusée dans des systèmes d'information. Hatchuel et Weil (1995) ont cependant évoqué les limites des systèmes à base de faits et de règles qui distinguent les connaissances et le raisonnement sur celles-ci. Cette distinction s'applique relativement bien aux savoirs qui suivent une « recette » mais moins bien à d'autres types d'apprentissages (ceux d'un stratège par exemple, qui prend en compte les relations organisationnelles que l'acteur entretient dans l'action).

Polanyi en 1966 définit deux catégories de connaissances dans les organisations (tacites et explicites) : « les connaissances explicites se réfèrent à la connaissance qui peut être exprimée sous forme de mots, de dessins, d'autres moyens "articulés" notamment les métaphores ; les connaissances tacites sont les connaissances qui sont difficilement exprimables quelle que soit la forme du langage » (Polanyi, 1966).

Néanmoins, il faudra attendre les travaux fondamentaux de Nonaka et Takeushi (1995) qui ont suggérés que la maitrise des modes de conversion explicite/tacite/ collectif/individuel est cruciale pour l'innovation. Le modèle de Nonaka et Takeuchi décrit le processus de création de connaissances comme un cycle avec des allers-retours répétitifs entre connaissances tacites et explicites, grâce à quatre formes de conversion (cf. figure 1.2)

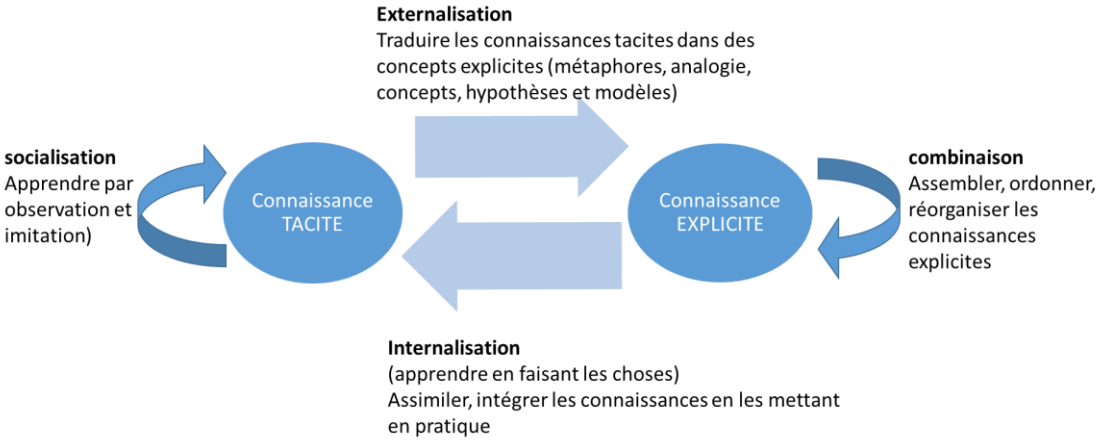


Figure 1.2 : Cycle de conversion tacite explicite (Nonaka et Takeushi, 1995)

Depuis les années 2000 il est admis, comme l'indique fort justement Hatchuel (2002) que « nous ne « recueillons » jamais vraiment la connaissance d'autrui, nous transformons plutôt la nôtre par des interactions avec lui. On ne peut donc traiter l'échange de connaissances comme la circulation d'une monnaie ». La connaissance s'incarne dans une personne et dans l'action (La connaissance est indissociable d'un sujet connaissant et d'un contexte. (Davenport, 1998)

En pratique, le processus général de construction d'une Mémoire d'Entreprise (ME) se décompose en plusieurs phases (Dieng-Kuntz, 2002) :

- Détection des besoins en mémoire d'entreprise
- Construction d'une mémoire d'entreprise
- Diffusion et utilisation d'une mémoire d'entreprise
- Évaluation et évolution de la mémoire d'entreprise

Là où le système expert espère une résolution, une mémoire d'entreprise vise principalement à aider l'utilisateur en lui fournissant des informations utiles. Les moyens employés peuvent être basés sur une mémoire documentaire, un système à base de connaissance ou à base de cas. Les méthodes comme CommonKads (Breuker, 1994) ou encore les ontologies (Guiarino, 2002) peuvent aider à la construction de la ME.

En ce qui concerne la diffusion des connaissances dans une organisation, on retrouve souvent les outils suivants :

- Système de gestion de contenu (structure, «taggage », arborescences, gestion du versionning (cycle de vie, publication)
- Gestion documentaire (Archivage et Gestion Electronique des Données)
- Espace de travail collaboratif (Intranet, worklow et portail de type Microsoft SharePoint)
- Communauté de pratiques (groupes de personnes qui partagent une identité et une pratique communes et qui, pour atteindre un objectif commun, échangent des connaissances) étayées par des réseaux sociaux entreprise (RSE) ou forums

Dans cette étude, de manière analogue aux méthodes en ingénierie des connaissances, nous allons procéder à l'explicitation de certains types de connaissance (comme le précise Nonaka et Takeushi dans leur modèle) afin de promouvoir l'apprentissage dans l'organisation. Nous adhérons à la définition de Hatchuel (1995) pour qui la connaissance née de l'interaction entre les acteurs. Aussi c'est dans ce sens que nous étudions les interactions professionnelles à travers les e-mails émis dans des projets.

Toutes ces techniques et méthodes sont établies à l'échelle de l'entreprise, mais également du projet. Comme l'objet de notre étude concerne les projets, nous allons rappeler quelques généralités sur les projets et leur gestion avant de détailler la mémoire de projet.

1.2 Qu'est qu'un Projet

En France, le terme «projet» à une définition normalisée AFNOR (voir aussi norme ISO 10006) :

- Afnor X50-115 : Un projet est un ensemble d'activités coordonnées et maîtrisées comportant des dates de début et de fin, entrepris dans le but d'atteindre un objectif conforme à des exigences spécifiques.

On notera que anglo-saxons englobent en général sous le terme projet aussi les acteurs, l'anticipation des risques, la planification. En effet, initialement, le concept de projet a été plutôt utilisé en ingénierie pour décrire la réalisation d'un système physique comme un nouveau produit industriel ou un nouveau bâtiment.

La raison essentielle qui pousse les entreprises à faire des projets est économique. Afin d'optimiser le ratio valeur/cout, la diminution des couts atteint rapidement ses limites (avec la concurrence et la mondialisation) et peut avoir un impact sur la qualité. Aussi pour augmenter la valeur, il est nécessaire d'innover. Le projet est l'unité de base dans une organisation pour développer les processus de changement.

Il faut en effet bien différencier les opérations normales de la vie d'une entreprise de divers projets qu'elle va entreprendre. Dans le mode opérationnel, l'entreprise effectue des tâches répétitives, identifiées, avec des acteurs connus, elle dégage un bénéfice, et doit amener une réponse immédiate en cas de problème (sinon on risque un blocage de la production). En mode projet, les tâches sont nouvelles, inconnues, les équipes temporaires, et il faut investir avant d'obtenir un résultat.

Un projet est défini et mis en œuvre pour élaborer la réponse au besoin d'un utilisateur, d'un client ou d'une clientèle. Il implique un objectif unique et mesurable et un ensemble d'actions ou de travaux qui concourent tous à sa réalisation. Par conséquent, la description d'un projet implique une définition claire des objectifs à atteindre qui doivent être traduits en buts. Un choix de technologies, des personnes et des ressources pour les atteindre, une planification et de la gestion de projet. Enfin un suivi, des évaluations et du contrôle sont indispensable tout au long de son déroulement.

1.3 Les éléments d'un projet

On peut considérer que les projets ont six caractéristiques essentielles:

1) Un projet est une opération temporaire ayant un début et une fin explicite: Lorsque les objectifs du projet ont été atteints ou lorsque le projet est terminé

parce que ses objectifs ne seront pas ou ne pourront pas être réalisés, la fin est réputée atteinte.

2) Le produit ou le service résultant d'un projet a un caractère unique et singulier
3) Un projet est novateur, unique et porteur de créativité. Il apporte une réponse novatrice à un besoin (problème) dans un contexte spécifique. Il est innovant et comporte donc une part d'incertitude liée à son caractère unique mais aussi au contexte parfois difficile à appréhender.
4) Une organisation temporaire : un ensemble de personnes et de ressources rassemblées de manière non permanente pour réaliser les tâches nécessaires à l'atteinte de l'objectif. Les projets sont complexes et impliquent des acteurs et des compétences variées coopérant ensemble.
5) Un résultat souhaité : des objectifs clairement définis et répondant à un besoin spécifique. Chaque projet crée un produit unique, service ou un résultat (parfois nommé livrable). Le résultat du projet peut être tangible (un produit, un prototype, un composant ou une amélioration d'un article existant) ou immatériel (un service, une amélioration d'un processus, un résultat de recherche parfois sous la forme d'un document).
6) Des contraintes : des ressources limitées, des délais ou des coûts.

Afin d'atteindre son objectif on met en place de la gestion de projet qui permet de mieux organiser/ répartir les tâches, et de s'assurer que les délais, les couts et la qualité requise seront respectés. Cet aspect-là est très intéressant pour notre étude. En effet du fait de la pluralité des acteurs (et de leur disciplines), des interactions seront obligatoires avec des objectifs communs. C'est donc un lieu de co-construction de connaissances nouvelles, avec une structure bien définie. Nous allons revoir les éléments principaux de la gestion d'un projet car elle permet de bien établir l'organisation du projet qui est très utile pour l'observation de la connaissance.

1.4 La Gestion de Projet

La gestion de projet consiste en la conception, la structuration, la planification, l'organisation, et le contrôle des projets. L'équipe d'un projet comprend donc le gestionnaire de projet (Project manager) et le groupe de personnes qui agissent de concert dans l'exécution des travaux du projet afin d'atteindre ses objectifs. Le Project manager est donc le garant des livrables ; il utilise une méthodologie scientifique et technique appropriée afin de remplir sa mission. Nous nous intéressons à la gestion de projet et non pas seulement à ses résultats, car elle nous permet de comprendre la dynamique de la création des connaissances à travers l'évolution du projet dans le temps.

1.4.1 Equipe d'un Projet

Nous présentons l'équipe, puisque comme nous l'avons stipulé précédemment, la connaissance est produite par les acteurs dans une organisation. Il faut alors comprendre le rôle de ses acteurs dans une organisation de projet. L'équipe de projet comprend donc le Project manager (et ses adjoints éventuels) et les autres membres de l'équipe qui effectuent le travail, mais qui ne sont pas nécessairement impliquées dans la gestion du projet.

1) Le rôle du chef du projet « Project manager » consiste en de la planification et de la gestion : des budgets, du reporting, du contrôle, des communications, des risques et de la partie administrative. Il peut être accompagné ponctuellement d'experts (pour la gestion financière, la logistique, le juridique/les contrats, l'ingénierie, la sécurité, les tests ou contrôles qualité)
2) Le rôle des membres de l'équipe projet : les membres de l'équipe qui effectuent le travail de création des livrables du projet. (Avec des connaissances et compétences spécifiques pour effectuer les tâches nécessaires aux projets)

On trouve aussi parfois :

1) Des représentants des clients ou utilisateurs. Ce sont les représentants des personnes (ou de l'entreprise) qui accepteront les livrables/produits du projet.

Ils peuvent également conseiller sur les exigences, assurer une bonne coordination ou valider l'acceptabilité des résultats du projet.

2) Les membres partenaires. Les membres des entreprises peuvent être affectés en tant que membres de l'équipe de projet afin d'assurer une bonne coordination.

Ces acteurs interagissent selon une planification mettant en avant plusieurs principales étapes comme (d'après Bachelet, 2013):

1) Rédaction et validation par les parties du Cahier des charges (CdC) fonctionnel (définition de l'objectif à atteindre non en termes de solutions mais de besoin.
2) A partir du CdC et des opérations nécessaires à la réalisation, on obtient le Diagramme des travaux
3) Diagramme des travaux avec distribution affectation des personnes = diagramme des responsabilités
4) Diagramme des responsabilités + durée et ordre des tâches (PERT méthodologie pour décrire des réseaux de tâches)
5) PERT + ressources disponibles = Gantt, calendrier du projet

En France, on retrouve souvent le mode organisationnel basé sur la maitrise d'œuvre et d'ouvrage :

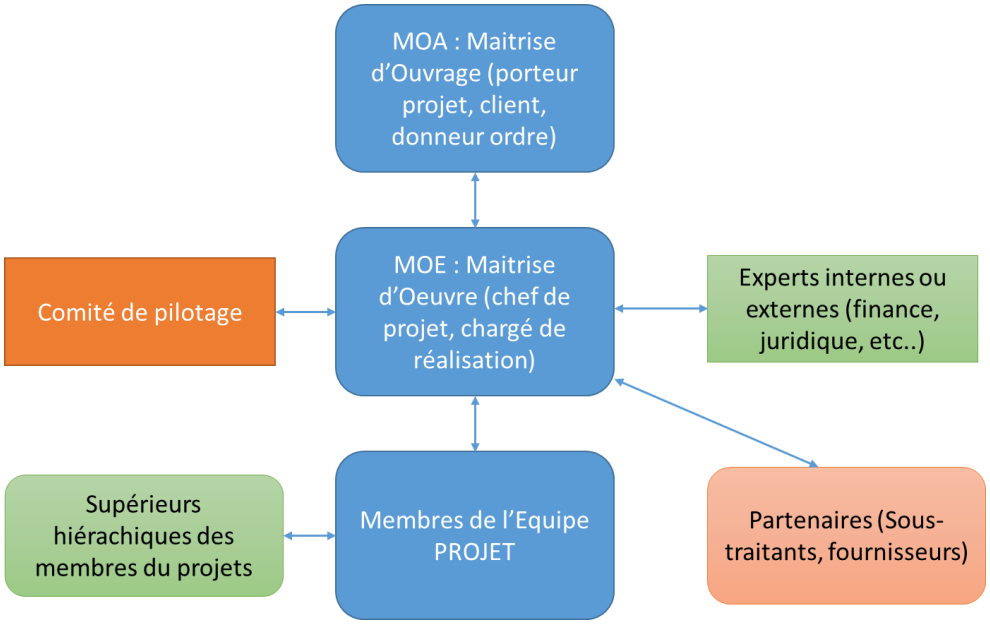


Figure 1.3 : Interaction MOA, MOE et équipe projet

La maitrise d'ouvrage (MOA) est responsable de l'expression du besoin (Cdc fonctionnel). Elle est à la base du projet, elle possède la connaissance métier nécessaire à la réalisation et fournit à la MOE les éléments prévus. Elle est en relation directe avec les futurs utilisateurs dont elle connaît parfaitement les méthodes de travail : les grandes fonctionnalités, les principales règles de gestion, valide et recette l'ensemble des livrables intermédiaires, et le produit final. Elle doit aussi normalement accompagner la conduite du changement auprès des utilisateurs.

La maitrise d'œuvre (MOE) assure la planification et la coordination de l'équipe de réalisation.

Elle définit et lance les études et développements spécifiés dans le contrat, le cahier des charges du projet. Elle assure la livraison et assiste la MOA ou le Client dans la recette et la mise en service du projet.

1.4.2 Phases et Cycle de vie d'un Projet

Un cycle de vie du projet est la série de phases qu'un projet traverse pendant son existence. La décomposition en phase poursuit des objectifs fonctionnels ou

techniques, ou se fait selon les résultats intermédiaires ou livrables, des jalons (Milestone) définis dans la politique globale de l'entreprise ou simplement des contraintes financières.

Les phases sont généralement bornées dans le temps, avec un début et de fin ou un point de contrôle. Les phases du projet sont employées lorsque la nature du travail à effectuer est unique à une partie du projet, et sont généralement liés à l'élaboration d'un livrable important. Le découpage en phase permet au projet d'être segmenté en sous-ensembles logiques pour la facilité de gestion, la planification, et le contrôle. Le travail effectué dans chaque phase est en général de nature différente, par conséquent les compétences requises de l'équipe de projet peuvent varier d'une phase à l'autre.

A un niveau global un projet se décomposera en une phase de définition du projet, une phase de réalisation du projet, et enfin une phase de capitalisation (bilan sur savoir-faire et expérience acquise sur laquelle nous reviendrons en détail plus loin).

La manière d'ordonnancer ces différentes phases définit le cycle de vie d'un projet. Le cycle de vie fournit le cadre de base pour la gestion du projet, quel que soit le travail spécifique impliqué.

La façon d'organiser le cycle de vie d'un projet peut varier dans un continuum de très prédictif à adaptatif. On peut signaler : Le cycle en cascade, le plus ancien : chaque phase attend la fin la suivante et les dates sont prédéfinies à l'avance. Le cycle en V, qui limite le retour aux étapes précédentes, amène les tests unitaires, d'intégrations (en informatique). Enfin le cycle de vie adaptatif est destiné à répondre à des changements rapides et à l'implication continue des intervenants. Il est également itératif (itération assez fréquentes de l'ordre de 2 à 4 semaines) et incrémental. Nous signalons l'accent fort mis sur les interactions entre les acteurs à travers ces cycles, notamment dans le cycle interactif. L'objectif projet sera décomposé en un ensemble d'exigences et de travaux à exécuter. Au début d'une itération, l'équipe travaillera à déterminer les éléments prioritaires sur la liste du carnet de commandes (parfois appelé un backlog de produit) pouvant être livrés dans la prochaine itération. A la fin de chaque itération, le produit devra être prêt pour une revue par le client. Les méthodes agiles (comme Scrum (Schwaber, 2004), XP (eXtreme programming) (Beck 2000) et la méthode RAD (Rapid Application Development) (Martin, 1991)

permettent de mettre en place ce type d'approche. Elles favorisent le travail collaboratif et augmentent la satisfaction client.

1.4.3 Gestion de Projet et communication

Les gestionnaires de projet passent la majeure partie de leur temps à communiquer avec les membres de l'équipe et d'autres parties prenantes du projet, qu'ils soient internes ou externe à l'organisation. Le succès d'un projet dans une entreprise est fortement dépendant d'une communication efficace entre les différents acteurs (elle facilite la prise de décision) à travers de laquelle de nouvelles connaissances sont construites. On peut distinguer différents canaux :

- la communication au sein de l'équipe
- la communication entre le chef de projet et les équipiers
- la communication MOE/MOA

La structure organisationnelle (matricielle, hiérarchique/fonctionnelle, en réseau) est un facteur environnemental de l'entreprise qui peut affecter la disponibilité des ressources et influencer la façon dont les projets sont menés. Elle aura un impact sur la communication également.

En plus des réunions en face à face et du téléphone, les intervenants et les membres de l'équipe de projet utilisent les communications électroniques (e-mail, messagerie instantanée, réseaux sociaux, SMS, web conférence, et d'autres formes de médias électroniques) pour communiquer de manière formelle ou informelle.

Une autre manière indirecte de communiquer s'effectue via les bases de connaissances de l'entreprise. Les bases contenant les normes, les politiques, les procédures (par exemple contrôle et qualité) de l'organisation, mais aussi des bases de données répertoriant les journées/hommes, les budgets.

On trouve parfois des données historiques et les « leçons » tirées de projets achevés. Cependant il faut souligner, et c'est tout l'intérêt de notre étude que ces informations

sont bien souvent limitées à des données quantitative sur la gestion des projets antérieurs (sur la mesure du rendement de projet précédents, les risques, les coûts, le calendrier, les diagrammes des tâches de l'échéancier du projet, …)

Enfin un mode de communication s'effectue via la gestion du projet elle-même et les traces qu'elle laisse dans le système de gestion de projet. Le Project manager va collecter, stocker, distribuer et/ou mettre à disposition les éléments clefs de la vie du projet en cours. Ce processus de suivi et de contrôle est essentiel à la bonne marche du projet, on y trouvera tous les documents de spécifications, les rapports, procès-verbaux de livraison, compte rendu de réunions, memo formels, rapport de performance, rendement, communication des progrès vers l'objectif, etc. Mais ces documents ne sont pas suffisants pour refléter la connaissance produite dans le projet. Une mise en correspondance de ces documents avec les interactions et le rôle des acteurs est primordiale à ce propos.

Dans le cadre de notre présente étude, il faut noter que certaines informations cruciales ne sont pas toujours préservées bien que faisant partie des communications. Lorsque le projet est terminé, on organise souvent une dernière réunion d'équipe qui est l'occasion d'un un bilan (appelée aussi réunion de post-mortem). Lors de cette réunion, les informations recueillies sont bien souvent les opinions des participants sur le déroulement du projet (questionnaire de type « êtes-vous contents du résultat du projet, de la manière de l'atteindre, quelle a été la partie la plus problématique, comment améliorerez-vous les choses à l'avenir, lequel de nos processus de travail à le mieux fonctionné, le moins bien ? » etc...)

Néanmoins les informations échangées lors des communications médiatisées entre les membres du projet resteront archivées dans leur boite email ou système de messagerie. C'est fort dommageable car elles contiennent souvent les causes des problèmes rencontrés, les raisonnements derrière les actions correctives choisies, et d'autres types de leçons apprises.

C'est tout l'objet de notre approche qui rejoint les problématiques de la Mémoire de projet (essentiellement en fin de celui-ci) et de la phase de capitalisation des connaissances. Les enseignements tirés de l'exécution d'un projet devraient être répertoriés et distribués de sorte qu'ils deviennent une partie de la base de données historique à la fois pour le projet et l'organisation du travail dans l'entreprise. Dans la partie suivante nous allons examiner ces aspects plus en détail avec la présentation de la Mémoire de Projet.

1.5 Mémoire de Projet

1.5.1 Présentation

Les projets de conception impliquent plusieurs acteurs de différents domaines pendant une période délimitée. Nous avons vu qu'un projet peut regrouper divers types d'intervenant provenant de domaine différents (section 1.4.1) et que ces équipes sont par nature éphémères. Or ces acteurs produisent des connaissances lors de l'interaction et prennent des décisions collaboratives. Donc, il est important de traiter également ce type de connaissance qui est généralement volatile (dissolution de l'équipe et mouvement de personnel). L'approche de ce type de connaissances, appelée Mémoire de Projet [Matta et al, 2000] doit représenter la dimension organisationnelle et coopérative de la connaissance. Les méthodes actuelles utilisées dans la gestion de la connaissance, fondée sur des interviews d'experts ne sont pas adaptées pour extraire ces dimensions de la connaissance. Pour faire face à la connaissance produite en activité de collaboration, nous avons besoin de techniques qui aident à extraire des connaissances du travail quotidien.

Actuellement, les concepteurs utilisent les connaissances tirées des projets passés, afin de mieux faire face aux projets à venir. Ils réutilisent la mémoire de la logique de conception pour faire face à de nouveaux problèmes et éviter les erreurs du passé. L'objectif est l'amélioration continue dans la manière de conduire les projets.

D'après Pomian (1996) la Mémoire de Projet est, par rapport à la Mémoire d'Entreprise, un volet restreint d'un exercice beaucoup plus large de capitalisation de toute une panoplie d'expériences réalisées au sein de l'entreprise par l'ensemble des gestionnaires, voire l'ensemble du personnel. Elle est cependant plus fréquente que

la Mémoire d'Entreprise en raison du fort turn-over et de l'aspect innovant des projets (particulièrement dans les projets informatiques). Elle contribuerait également à l'amélioration de l'ensemble de l'organisation (évolution des normes, des procédures, des moyens de communication, mais aussi une meilleure gestion des ressources humaines (besoin de formation, évaluation et valorisations des compétences, affectation des ressources)).

Elle n'est pas seulement tournée vers l'avenir et si elle est appliquée pendant le déroulement d'un projet (nous reviendrons sur cet aspect dans la partie), elle permettrait un meilleur contrôle et des réajustements plus rapides.

1.5.2 Définition

Dans (Matta, 2000), (Dieng-Kuntz, 1998), les auteurs définissent la Mémoire de Projet comme une « mémoire des connaissances et des informations, acquises et produites au cours de la réalisation des projets ». Elle doit principalement tenir compte de (voir figure 1.4):

- L'organisation du projet: les différents participants, leurs compétences, leur organisation en équipes, les tâches qui sont assignées à chaque participant, etc.
- Les cadres de référence (règles, méthodes, lois, ...) utilisés dans les différentes étapes du projet.
- La réalisation du projet: la résolution de problèmes potentiels, l'évaluation des solutions ainsi que la gestion des incidents qui se sont produit.
- Le processus de prise de décision: la stratégie de négociation, qui guide la prise de décisions ainsi que les résultats des décisions.

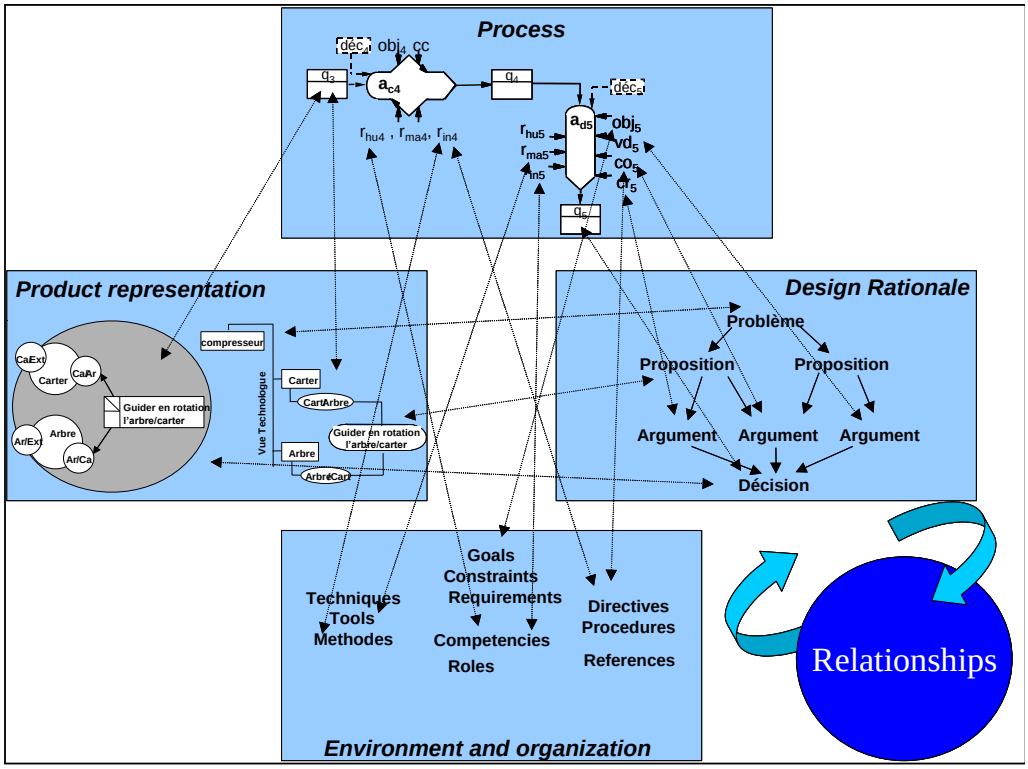


Figure 1.4 : Mémoire de Projet

Souvent, il y a des relations d'interdépendance entre les différents éléments d'une Mémoire de Projet. Grâce à l'analyse de ces relations, il est possible de rendre pertinent et explicite les connaissances utilisées dans la réalisation du projet. La traçabilité de ce type de mémoire peut être guidée par des études sur la logique de conception et par des techniques d'ingénierie des connaissances. Cette mémoire doit contenir des éléments d'expérience à la fois du contexte de la résolution de problèmes et de sa résolution.

1.5.3 Mise en Œuvre

La mise en place d'une Mémoire de Projet comme tous les projet de Knowledge Management appelle à des choix : Que documenter, qui doit faire le recueil, à quel moment et comment, et enfin quelle est la meilleure manière de l'exploiter par la suite.

QUOI

En général et comme mentionné plus haut, on capitalisera :

- La définition du projet, l'historique et le résultat (objectifs initiaux et réalisés, équipe projets et moyens, étapes, livrables)
- Les processus : documenter le processus de prise de décisions déterminantes, les essais, les erreurs, mais également le contexte (c'est cela qui permettra d'augmenter la capacité de généralisation inter projet de de la MP). Un projet est un processus dynamique qui se développe en contexte et non, par exemple, « une banque de connaissances sur ce que les gens savent à la fin d'un projet » (Pomian 1996). Au niveau de la diffusion, cette approche nécessite un choix d'outils (livres, système informatique, trame narrative, etc..) plus pratique à mettre en œuvre que des normes ou des procédures. Enfin il faut remarquer que les membres de l'équipe projet auront une tendance naturelle à ne pas mentionner les échecs et les voies sans issues qu'ils ont explorées.

QUAND

La tendance naturelle est de la constituer à l'achèvement du projet, une fois que les résultats sont obtenus. Mais cela fait courir le risque d'omettre des étapes importantes à cause du facteur oubli et de la connaissance du résultat final. Se concentrer seulement sur la réunion Post Mortem focaliserait la MP sur les résultats, et sur des aspects de gestion de projet (écarts par rapport aux prévisions, retards, coût). Constituer une Mémoire en cours de projet est délicat : comment décider de la pertinence d'un résultat ou d'un processus intérimaire ? Et surtout, il ne faut pas rajouter de travail supplémentaire à l'équipe projet dont la mission essentielle est la réussite de celui-ci. Nous reviendrons plus loin sur des approches moins intrusives pour réussir cet exercice en détournant les outils usuels de l'équipe projet.

PAR QUI

Les membres de l'équipe projet par exemple via un journal de bord ou questionnaire (Pomian, 1996), mais ceci ne sera pas approprié pour le processus de décision, de résolution de problème et toute connaissance née d'activité collaborative. En cas de

savoir-faire précis, Des intervenants extérieurs sont nécessaires car du fait des connaissances tacites, il y aurait un écart entre ce que les gens disent à propos de ce qu'ils font et ce qu'ils font réellement, écart qui nécessiterait un recueil d'information médiatisées (Pomian, 1996). Enfin il faudrait idéalement conserver également une trace des erreurs et mauvais choix de l'équipe (que les utilisateurs seront moins enclins à divulguer pour se concentrer davantage sur le résultat final).

1.6 La Logique de Conception

Revenons en particulier sur place de la logique de conception (ou Design Rationale) : si la Mémoire de Projet se focalise effectivement sur la conservation de « la définition du projet, les activités, l'historique et le résultat » (Tourtier, 1995). La résolution de problèmes (ou Problem Solving) en est une composante essentielle car elle aborde la définition du problème, les suggestions, et les choix et décisions. A ce titre, elle fait partie de la mémoire de logique de conception (Matta, 2000).

Certaines méthodes d'ingénierie des connaissances visent davantage à la capitalisation des connaissances (REX, MEREX, MASK) d'autres sont plus orientées sur la traçabilité du Design Rationale. Conserver une mémoire des résultats seuls ne suffit pas, il faut également une mémoire des processus décisionnels et de leur contexte, plus susceptible d'apporter une valeur ajoutée à l'entreprise pour l'apprentissage pour de futurs projets. Ces deux types de mémoire (résultats/processus) sont complémentaires.

C'est essentiellement lors des réunions et discussions, ou des conversations téléphoniques que sont présentés les problèmes de réalisation, les solutions possibles et qu'il y a argumentation et négociation. Un ensemble de méthodes (ex : IBIS, QOC, DRCS, DIPA (Matta, 2000)) ont pour objectif de modéliser et de représenter les choix effectués lors d'un projet de conception ainsi que les justifications de ces choix. Elles permettent de justifier les résultats obtenus, ainsi que les solutions écartées et leurs conséquences.

La logique de conception représente l'ensemble des problèmes, solutions, décisions et informations relatives au déroulement d'un projet (Buckigham Shum, 1997). Ces informations doivent couvrir l'ensemble des étapes d'un projet. L'intérêt de

capitaliser ces connaissances est de favoriser la contextualisation. Lors d'un projet (et surtout pendant la conception) de la connaissance est créée ou manipulée dans un contexte spécifique. En effet, les projets de conception forment une composante cruciale de l'entreprise. Ils portent sur le développement de nouveaux produits qui vont constituer l'avantage concurrentiel. De nombreux acteurs avec des compétences variées participent aux prises de décision dans les différentes phases de la conception, ce qui place la connaissance au centre de ce processus. Il est alors primordial de capitaliser dans les mémoires de projet de conception, les différents points de vue des acteurs dans le processus de décision.

Pour résumer de manière simplifiée le processus de conception, on peut le voir comme la prise en compte des contraintes issues du problème (formulé à partir de l'expression du besoin). Des solutions correspondantes sont alors proposées et évaluées itérativement (face à des alternatives pour résoudre un problème, on expérimente, puis on recommence) jusqu'à satisfaction des contraintes. Les concepteurs au cours de leur activité doivent prendre des décisions. Cet historique de succès et d'échecs, la faculté de faire les bons choix constitue leur expérience mais elle reste souvent personnelle et tacite.

Une représentation claire du contexte et de la logique de conception peut être trouvée dans (Bekhti et al, 2003) et est présentée dans la figure 1.5. Ce contexte peut inclure les décisions de choix ou de rejets, les arguments, les objections, etc.

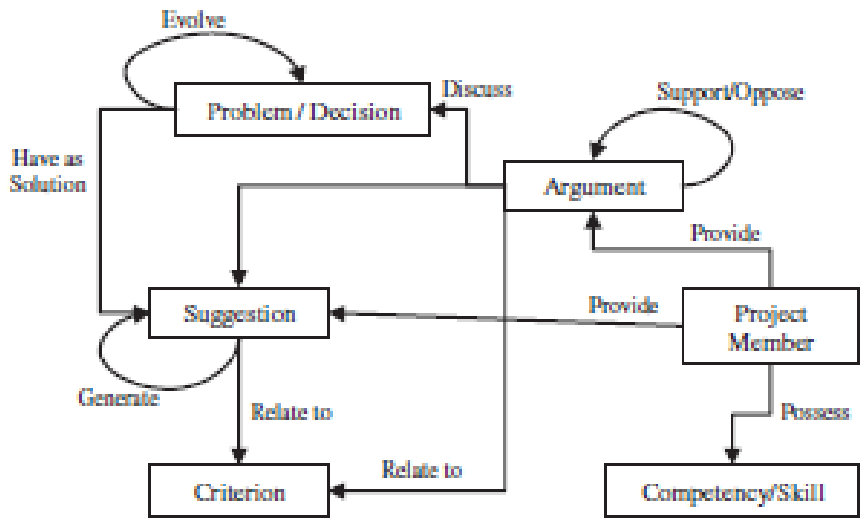


Figure 1.5 : La résolution de problèmes et la logique de conception dans leur contexte

1.7 Récapitulatif sur les modèles de connaissances et les projets

Dans les parties 1.1 à 1.4, nous avons présenté brièvement le KM et la gestion de projet. Enfin dans les partie 1.5 et 1.6 nous avons examiné plus en détail la Mémoire de Projet et la logique de conception.

- La gestion des connaissances est une démarche indispensable à l'entreprise actuelle. En particulier dans cadre de projet qui impliquent de l'innovation, des équipes temporaires avec des compétences variées, et donc un risque important de disparition de la connaissance collaborative qui a été produite.

- La Mémoire de projet apporte des réponses dans cette optique. Surtout en ce qui concerne la mémoire de logique de conception qui vise à capturer en contexte la démarche de résolution de problème de l'équipe projet.

Cependant et c'est l'objet de notre étude sur la traçabilité en Mémoire de projet, quelques aspects restent encore à améliorer :

- La capture/ le recueil :

Nous avons noté qu'il est difficile d'apprendre a posteriori en interrogeant les participants, cela pose des problèmes de sélection de ce qui est important, à capitaliser ou non et d'oubli. Et apprendre tout au long du projet est également délicat sans rajouter du travail aux équipes avec le même souci de pertinence des informations à retenir avant l'achèvement complet du projet

- L'exploitation :

Comme souvent en KM et encore davantage en mémoire de Projet, se pose la question de l'exploitation des leçons apprises. L'accès aux informations stockées sera problématique si par exemple dans un réseau d'entreprise on consacre un répertoire à chaque projet, comment feront les utilisateurs futurs pour retrouver des leçons apprises étalées dans des documents différents sur des dizaines de répertoires ? Et comment estimer leur impact sur leur projet actuel ?

Notre étude se place dans le cadre de la mémoire de projet, notre but est d'apprendre des projets passés. En outre nous avons vu au début que la connaissance se transmet difficilement (tacite vers explicite surtout quand collaborative). Et pourtant il est impératif de conserver la connaissance de la logique de conception. Ceci pour deux raison

- A la fois la réutilisation dans des projets similaires
- Mais aussi, et c'est un cas extrêmement fréquent en entreprise, une fois l'équipe projet dissoute, il arrive parfois que des questions surgissent sur le projet des mois voire des années après.

Les pistes que nous avons retenues à fin d'investigation sont celles de la traçabilité, et du processus de résolution de problème. En outre les moyens de communication électroniques (si les circonstances du projet s'y prêtent) pourraient être un moyen moins intrusif et moins gourmand en ressource pour recueillir des éléments de connaissance. (Il y a bien entendu une différence entre le processus de décision et les traces de celui-ci). Surtout que ces moyens sont support à l'interaction entre les acteurs, promoteur de la création des connaissances. Cependant ces traces ne sont pas inutiles : elles peuvent servir à illustrer et à pister les décisions prises en cours de projet (Pomian, 1996).

On voit donc se dessiner les grands questionnements et axes de notre recherche :

- Nous allons examiner dans quelle mesure et comment les communications médiatisées (dans un projet où elles sont très utilisées) peuvent participer à la

traçabilité de la Mémoire de Projet. Mais aussi de quelle manière aborder la capture, la structuration et explorer un moyen de diffusion

Nous présentons donc dans les chapitres suivant d'une part les théories sur lesquelles se basent notre étude et d'autre part, l'approche que nous avons développée afin de permettre une traçabilité et une structuration des connaissances à partir des communications médiatisées entre les acteurs d'un projet.

Chapitre 2. TRAÇABILITE ET RESOLUTION DE PROBLEME

2.1 La Traçabilité

Le rôle de la traçabilité a été reconnu dans des travaux pionniers des années 60 (Naur et Randell, 1969) à propos des problèmes de génie logiciel. Il fallait trouver une méthode efficace en matière de conception de systèmes informatiques afin de s'assurer que le système en cours d'élaboration contienne des traces explicite de sa conception.

Il faudra cependant attendre les travaux de (Ramesh et Edwards, 1993; Gotel et Finkelstein, 1994), sur les questions et les problèmes associés à la traçabilité pour avoir une analyse systématique.

En particulier Ramesh et al. ont détaillé les bénéfices de la traçabilité (Le but ultime de toute stratégie de traçabilité étant d'améliorer la performance des activités futures). Les utilisations potentielles des données de traçabilité seront par exemple pour les différents intervenants d'un projet :

Gestionnaire projet (MOE) :

- La réalisation d'analyses d'impact pour estimer l'effort de changement (facilitée lorsque les exigences sont reliées par les traces aux composants qui les implémentent).
- Les conflits dans les exigences peuvent être découverts plut tôt et éviter des retards.
- Le respect des processus et des politiques qualité, notamment pour les tests.
- Les exigences non satisfaites, recueillies et la charge nécessaire estimée pour les satisfaire.
- Les produits et services futurs pour la réutilisation des décisions d'implémentation des projets passés (d'où un gain de temps).
- L'identification des éléments potentiels pour la réutilisation.
- Enfin un système de traces procure au gestionnaire de projet un suivi plus exact de l'avancement du projet.

Client (MOA)

- Quelle partie d'un système doivent être testé pour vérifier les exigences (assurer la couverture de test nécessaire et suffisante).
- Chaque partie d'un système peut être liée par la trace à une exigence (afin d'être certain qu'il n'y a rien de superflu) et que le produit est conforme.

Concepteur/Equipe projet

- Le système de traçabilité devrait enregistrer les résultats de la conception, les justifications des décisions, les solutions envisagées et les hypothèses formulées. Si ceux-ci sont stockés avec les liens entre les exigences et la conception ceci permet de vérifier plus facilement que la conception satisfait aux exigences.
- Estimation de l'impact d'un changement dans les exigences relatives à la conception.
- Le concepteur peut comprendre les raisons pour lesquelles une certaine conception a été retenue et une autre rejetée même si la conception a été

produite il y a longtemps par un concepteur ne faisant plus partie de l'entreprise. (Reconstruction des décisions antérieures).

- Ces raisons peuvent concerner les décisions de conception à des exigences non fonctionnelles (par exemple d'un changement dans la technologie de mise en œuvre).
- Le concepteur peut réutiliser les composants de conception dans d'autres projets, car les hypothèses et le contexte dans lesquelles un composant ou une solution a fonctionné ont été enregistrés.
- Le concepteur peut estimer l'interdépendance des exigences.

En particulier et ceci est en relation directe avec notre étude : Ramesh et al. citent un exemple extrême d'un projet où l'entreprise a été forcée de réembaucher des ingénieurs qui avaient quitté la société afin de reconstruire les raisons derrière la logique de conception originale (Ramesh, 1995); Cette situation aurait pu être évitée si les raisons des décisions avaient été plus facilement traçables.

Certes la traçabilité a un impact non négligeable sur la vie quotidienne du projet et il y a donc un compromis à trouver entre l'augmentation des coûts due au recueil des données de traçabilité et la réduction ultérieure des coûts de réalisation à attendre.

Par ailleurs la traçabilité est toujours un problème d'actualité avec par exemple les études d'A. Mille (Mille, 2005) sur le raisonnement à base de traces, ou encore sur la génération automatique de traces (Cleland-Huang, 2007). La valeur d'un système de gestion des traces est essentiellement dépendante de la capacité à produire des traces pertinentes et exploitables de l'activité individuelle ou collective des intervenants d'un projet.

Mais il nous faut d'abord revenir sur les définitions usuelles de traces et de traçabilité qui seront l'objet des parties suivantes.

2.2 Trace et Traçabilité

2.2.1 Trace

La définition usuelle d'une Trace dans le cadre de la théorie de la traçabilité pour la conception d'un produit ou d'un service est la suivante :

Trace : Un triplé spécifié d'éléments comprenant: un artefact source, un artefact cible et une liaison (dite de trace) associant les deux objets. (si plus de deux objets, les objets agrégés sont considérés comme un seul artefact de trace). Et par extension le fait de suivre un lien de trace d'un artefact source à un artefact cible (traçabilité avant) ou vice-versa (traçabilité arrière). (Ramesh, 1995)

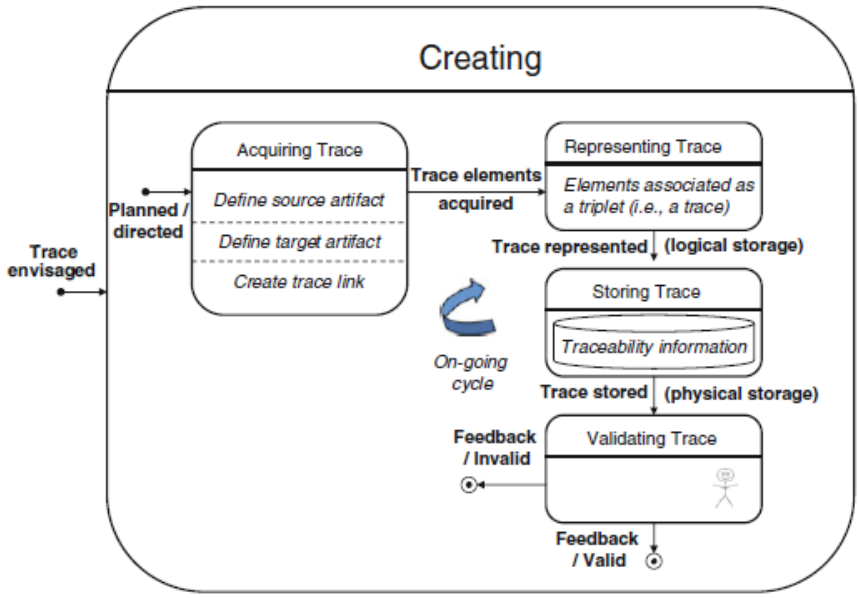


Figure 2.1 : Création d'une trace entre artefacts (d'après Cleland-Huang, 2012)

2.2.2 Traçabilité

À son tour, la traçabilité est simplement le « le potentiel de ces traces à être établi (créé et maintenu) et utilisé ». Cependant, elle dépend de la nature et de l'emplacement des artefacts à relier, des mécaniques qui vont créer et maintenir ces liens, et des types d'utilisation qui en seront fait.

D'autres définitions communes mettent en avant la capacité de mettre en œuvre et de conserver la trace d'un ensemble ou d'un type d'information donné avec une certaine

granularité, ou la capacité de mettre en relation chronologique des entités identifiables d'une manière qui soit vérifiable.

Le défi pour la traçabilité est que chacun des éléments constitutifs (expression des besoins, cahier des charges, artefacts au divers degré d'avancement) soit correctement recueilli, horodaté, représenté et stocké, et ensuite récupéré comme une trace pour étude et réutilisation.

En pratique, on peut distinguer différentes sorte de traçabilité. La plus répandue en matière de projet est la traçabilité des exigences. La traçabilité des exigences se concentre sur les liens entre les exigences et artefacts produits (Gotel et Finkelstein, 1994). Plus précisément Gotel et al. définissent la traçabilité des exigences comme « la faculté de décrire et de suivre le cycle de vie d'une exigence à la fois vers l'avant et l'arrière, c'est-à-dire depuis ses origines, vers les spécifications et le développement, puis le déploiement et l'utilisation, et sur les périodes d'évolution et d'itération de ces phases ». La partie exigence du cahier des charges décrit les objectifs du produit et le comportement observable que doit avoir le produit afin d'avoir atteint ces objectifs. Pour certains, la gestion des exigences englobe le passage de l'expression des besoins aux spécifications (Davis, 2005) alors que pour d'autres elle est distincte de celle-ci (Wieger, 1999).

Dans le développement de logiciels; la traçabilité des exigences se réfère à la capacité à lier les exigences du produit aux justifications des parties prenantes (équipe projet, Project manager et client) et à transmettre à des artefacts de conception, le code, et des cas de test correspondants. Il s'agit de suivre la vie d'une exigence à la fois dans les directions avant et arrière.

Comme nous l'avons évoqué précédemment, la traçabilité des exigences permet une analyse de l'impact des changements, la vérification de la conformité, la sélection des tests, et la validation des exigences. Elle est habituellement réalisée sous la forme d'une matrice créée pour la vérification et la validation du projet. Malheureusement, la construction et l'entretien d'une matrice de trace des exigences (MTR) peut être

très ardue. Aussi d'autres approches que nous verrons plus loin ont été développées pour la génération automatique de traces à l'aide de la recherche d'information.

2.3 Capture

Les projets actuels (et surtout en génie logiciel) sont souvent complexes et comprennent une multitude de produits intermédiaires et connexes, en plus du résultat final.

Par exemple un logiciel comprendra des objets tels que: le code source, l'expression des besoins, le cahier des charges fonctionnel, les documents de conception, les documents d'exigences, cas de test, rapports de bugs, les communications entre les parties prenantes, etc.

Ceux-ci sont créés et maintenus sur une longue période de temps par différentes personnes. Aussi il est difficile d'établir et de maintenir des liens explicites entre les artefacts logiciels. C'est également le cas lorsque le projet implique la création et la manipulation d'artefacts de nature hétérogène (produits physiques, procédures, logiciels, services, vidéos, etc.)

Or la capture des traces implique la création de liens en même temps que la création des artefacts qu'ils associent. Capturer manuellement les liens de traçabilité représente un travail fastidieux et très long et, sans aucun soutien automatisé, c'est une tâche quasi impossible.

Aussi dans la pratique, en particulier pour les projets de grande envergure, il est fait appel à des outils et des techniques pour récupérer la traçabilité. Grâce à eux les liens de traces peuvent alors être créés automatiquement ou semi automatiquement.

2.4 Techniques pour automatiser la Traçabilité

Dans le cas des systèmes informatique par exemple, un défi majeur dans la reprise des liens de traçabilité entre les artefacts de logiciels réside dans le fait que ces objets se

présentent sous des formats et des niveaux d'abstraction différents. De plus la sémantique de ces liens peut être interprétée différemment selon les personnes.

Cependant, il y a un type de données présent dans tous les artefacts logiciels: les données textuelles. L'extraction et l'analyse de ces données est essentielle à la découverte des liens de traçabilité via des outils appropriés.

La plupart du temps les parties textuelles décrivent les objets et leur sémantique. L'hypothèse est que si le contenu textuel de deux artefacts a trait à des concepts similaires, alors les deux objets sont conceptuellement liés et qu'un lien de traçabilité pourrait être établi.

IR

Une solution adoptée par les chercheurs et l'ingénierie des connaissances pour extraire et analyser les données textuelles intégrées dans artefacts logiciels est l'utilisation de la recherche d'information. (**IR – Information Retrieval**) (Baeza-Yates et Ribeiro-Neto, 1999).

Les méthodologies IR pour récupérer les liens de traçabilité se base sur la similitude entre le texte contenu dans les artefacts de produits. Plus la similitude textuelle entre deux artefacts est grande, plus il est probable qu'il existe un lien entre eux. L'avantage des techniques d'IR est qu'elles ne reposent pas sur un vocabulaire prédéfini ou une grammaire ce qui réduit considérablement les coûts de récupération des liens. (Andrea De Lucia, 2012). En outre ces méthodes peuvent être appliquées pendant le cycle de vie du projet ou après que le produit soit terminé.

Le processus de récupération de lien de traçabilité en utilisant des méthodes IR prendra la forme suivante:

- ⇨ Analyse des documents, prétraitements et extraction
- ⇨ Indexation de corpus avec une méthode d'IR (voir partie chapitre 4);
- ⇨ Génération de liste de classement;
- ⇨ Analyse des liens candidats pour ne retenir que les pertinents.

Dans le cadre du développement logiciel ces approches ont été appliquées (Andrea De Lucia 2012) aux artefacts logiciels suivants : les exigences et le code source (Antoniol, 2000), documentation externe et la documentation de conception (Capobianco, 2009) les cas de test et les rapports d'erreur reports (Yadla, 2005), et enfin les messages électroniques ou « emails » (Bacchelli, 2010).

Cleland-Huang et al (2004) suggèrent qu'une stratégie de traçabilité raisonnable est de « maximiser l'utilisation de la génération de liens dynamiques », via l'utilisation de l'IR ou d'heuristiques.

Nous retiendrons de ces méthodes qu'elles sont sans doute nécessaires dans le cadre de projets complexes où une ouverture manuelle et complète de tous les liens de traçabilité serait difficilement envisageable. Cependant deux points méritent d'être soulignés :

- ⇨ Cette reconstruction peut signaler des liens de traçabilité potentiels entre des artefacts projet mais pas la nature de la relation.
- ⇨ Le contexte humain, organisationnel n'est absolument pas pris en compte, ce sont des méthodes de Textmining basées essentiellement sur une analyse syntaxique.

Garder la trace de tous les artefacts de produit, des abstractions, des relations et des transformations tout au long du cycle de vie implique une gestion complexe. Cela rejoint une problématique plus large : comment tracer au-delà des simples exigences, comment tracer le processus décisionnel des membres de l'équipe de conception et en retirer des enseignements? Nous allons approfondir le domaine qui retient notre attention dans cette étude : la logique de conception et en particulier la résolution de problème. Nous allons nous attarder sur ces deux aspects dans leurs liens avec la traçabilité.

2.5 Traces et Raisonnement : Trace Based Reasoning

2.5.1 Généralité raisonnement connaissances trace

Tout raisonnement se fonde sur le rappel de connaissances antérieures afin d'atteindre un objectif. L'expérience étant la base de chaque connaissance, les raisonnements sont basés sur l'expérience (Richard, 1990).

En psychologie cognitive, la connaissance déclarative (explicitée avec des symboles) relève de l'expertise alors que la pratique (qui n'est pas déclarative) est plus implicite (Kirsner, 2013). Aussi les traces d'une pratique dans l'environnement ne sont que des enregistrements indirects d'une connaissance qui émerge de l'action concrète.

Nous avons vu que la capture de la connaissance en contexte est un processus délicat et souvent les enregistrements de connaissance peuvent être le résultat d'un processus explicite par des ingénieurs de la connaissance (Charlet, 2004). Une telle démarche vise à représenter la connaissance d'une manière qui soit exploitable par un environnement informatique. Ceci pour rendre l'émergence de connaissance plus facile durant une expérience similaire s'ils sont disponibles dans l'environnement

Le raisonnement à base de cas (Case Based Reasoning ou CBR) ambitionne aussi d'aider à résoudre des problèmes en réutilisant des problèmes déjà résolus (Kolodner, 1993). Il faut cependant structurer la connaissance et construire une bibliothèque de cas de problèmes. Puis lors de l'utilisation pratique, en face d'un problème recherché trouver un cas présentant des similitudes dans la base existante.

Avec A. Mille (Mille 2001), nous partageons l'idée que l'expérience humaine temporellement située par définition est bien représentée par un enregistrement temporel ou une trace décrivant les processus sous-jacents. Nous allons par la suite présenter les travaux d'A. Mille et de son équipe sur le **TBR** (**Trace Based Reasoning**).

2.5.2 Définition TBR

Le Raisonnement à Base de Trace (TBR), est un modèle de raisonnement fondé sur les traces d'interactions laissées par les utilisateurs dans les environnements numériques. Le TBR facilite la réutilisation de l'expérience car les traces d'interactions enregistrent les expériences des utilisateurs en situation de résolution de problème et en contexte. De plus, les traces d'interaction peuvent être utilisées comme une source de connaissance pour découvrir d'autres connaissances utiles pour le processus de raisonnement (Cordier et al, 2010)

L'idée de prendre en compte le contexte dans le processus de raisonnement a été aussi explorée. Par exemple Zimmerman (Zimmerman 2003) montre que la combinaison de la méthodologie CBR et du context awareness (la sensibilité au contexte, la perception du contexte) est une façon nouvelle et puissante de modéliser et de raisonner à partir des contextes

Une trace est communément définie comme un ensemble d'élément situé temporellement (par exemple les fichiers logs, un flux de tweets ou RSS). Sans modèle, ces traces sont difficiles à exploiter à moins d'un logiciel spécifique. L'approche TBR associe un modèle avec une trace, et il est possible de raisonner (par inférence) dessus.

Le modèle d'une trace (ou M-Trace) est la description formelle de la structure et du contenu de la trace. Il procure des informations sur :

- la propriété de la trace (unité de temps, longueur, etc..),
- les éléments de la trace (les « obsel » : éléments observables avec leurs attributs, et l'identifiant « id » de l'utilisateur, etc..)
- les relations entre ces éléments (y compris intervalle temporel, ordre, etc.)

Un « épisode » est un modèle de trace représentant une tâche ou une expérience donnée qui sera rappelé quand un besoin spécifique apparait. On peut donc retrouver et utiliser des éléments de contexte relatifs à l'épisode courant.

La signature d'un épisode est une spécification des contraintes et propriété distinctives caractéristiques (pattern=schéma, durée) que l'épisode doit satisfaire. Ceci permet de retrouver un épisode dans les traces. Il faut noter que contrairement au CBR, l'expérience n'est pas conservée en cas structurés. Par exemple dans le modèle CBR, il est possible de combiner l'expérience de plusieurs utilisateurs pour résoudre un problème avec un processus de raisonnement basé sur la fusion des traces individuelle.

En pratique, le TBR fonctionne ainsi : nous reprenons l'exemple de Mille (2001) sur l'assistance à l'utilisateur (le but est de retrouver une situation passée similaire). Une signature de l'épisode courant est créée à partir de fragment de la trace actuelle de l'utilisateur. Un ensemble de mesures de similarité est associé à cet épisode qui permet de retrouver l'épisode le plus similaire dans la trace. Les traces sont exploitées sur la base de similarité de patterns.

2.5.3 Différences entre CBR et TBR en regard de l'analyse des emails

Même si poursuivant des objectifs similaires, TBR et CBR diffèrent au moins sur les points suivants:

- dans leur méthode de stockage (connaissance structurée en cas et définie à la conception en CBR)
- gestion de la temporalité (peu d'horodatage dans les descripteurs de cas)
- prise en compte du contexte, (en CBR, le système recherche un cas similaire au cas actuel mais sans la partie solution, en TBR cela dépendra de l'épisode (tâche), cela aura aussi un impact sur l'adaptation lors de la présentation à l'utilisateur dans le cas d'une réutilisation)
- incorporation des nouveautés au système (en CBR cela dépendra de la pertinence du cas, alors qu'en TBR les traces se rajoutent)

2.5.4 TBR et résolution de problème

Les traces gardent un enregistrement des interactions, et donc du processus de résolution de problème. Les traces d'interaction avec l'environnement peuvent contenir des connaissances dans lesquelles les expériences de résolution de problème sont stockées implicitement et en contexte (mais à la différence de CBR, elle n'est pas structurée). Plus proche de notre problématique, TBR s'est révélé un outil efficace par exemple dans le cadre de l'assistance aux utilisateurs. Cordier et al. (Cordier, Mascret, et Mille, 2010) montre que :

- Les traces sont des objets réflexifs (pour obtenir une réflexivité sur l'activité dans le cadre d'activité médiées complexes, une représentation de l'activité est nécessaire). Zarka et al. ont montré que le simple fait de rejouer une trace fournit une forme d'assistance efficace (Zarka et al.2011)
- Les traces peuvent être partagées entre les utilisateurs ce qui facilite le partage d'expérience
- Les traces peuvent être transformées ce qui les rend utilisables à différent niveau et processus
- Les traces sont de bons réceptacles de connaissance. Si elles sont structurées pour faciliter leur réutilisation, elles permettent le recueil, la gestion et la restitution des connaissances aux utilisateurs. Les documents projets constituent également des traces mais ne sont pas toujours faciles à réutiliser.

2.6 Traçabilité et Logique de conception

Nous avons vu dans le chapitre 1 que la logique de conception « Design rationale » ne peut pas être obtenu à partir des spécifications de conception car il n'y pas de pratique systématique pour la capturer et que même si c'est partiellement le cas, elle n'est pas organisée de telle manière qu'elle puisse être retrouvé et tracée efficacement.

Or la logique de conception conserve des connaissances de conception (les hypothèses/contraintes, le raisonnement, des alternatives parfois non retenues, le

contexte environnemental (par exemple technologies disponibles à une date)). La représentation basée sur l'argumentation utilise des nœuds et des liens pour représenter la connaissance et les relations

Des méthodes de représentation de la logique de conception basées sur l'argumentation représentent les délibérations ayant lieu pendant la conception (Conklin et Begeman, 1988). Les modèles initiaux avec des nœuds (connaissance) et des liens (relations) datent de Toulmin (Toulmin, 1958) et depuis ont évolué vers une représentation des relations entre les questions, les propositions et les arguments comme les méthodes IBIS « Issue-Based Information System » (Conklin et Begeman, 1988) et QOC (Shum, 1997), mais ne fournissent pas de moyens pour recueillir la logique de conception ou la communiquer.

Or il est critique de pouvoir capturer les raisons derrière les décisions de conception pour l'évolution ou la maintenance du produit. Pendant le processus de conception, des décisions sont prises mais souvent les justifications ne sont souvent pas enregistrées (Parnas et Clements, 1985) Quand on leur demande d'expliquer leur choix les concepteurs doivent soit se rappeler les raisons ou les reconstruire (Gruber and Russell, 1996). Ceci est dû au fait que les méthodes actuelles ne mettent pas en relation les objets de conception avec le raisonnement de conception (Herbsleb et Kuwana, 1993).

Un système efficace devrait être capable de récupérer le design rationale en lien avec une explication claire des objets de conception. Ce système devra prendre en compte l'identification des informations requises, modéliser la logique de conception et son lien avec les objets conçus, et enfin utiliser ce modèle pour la traçabilité du raisonnement de conception.

La traçabilité fournit en général un moyen de trouver un lien entre les artefacts de conception (par exemple les exigences) et les objets de conception. Néanmoins un élément clef est manquant pour améliorer la vérification de la conformité et l'évolution d'un produit : la logique de conception qui contient les hypothèses implicite, les contraintes, le raisonnement. Des outils supportant la traçabilité et

capturant le design rationale dont nécessaires : ils devraient relier par des traces les objets conçus et les exigences à la logique de conception. La méthode DYPKM (Bekhti, 2003), (Matta, 2014) permet de recueillir la logique de décision à partir des réunions en procédant par des enregistrements structurés de la prise de décision. Cependant, la logique de conception est également présente à travers les interactions médiatisées, que ces approches ne permettent pas d'explorer.

2.6.1 Importance du recueil de la Logique de conception

Les raisons suivantes d'après Tang et al. (Tang 2007) expliquent pourquoi la traçabilité de la logique de conception pourrait aider dans le cas du cycle de vie d'un développement d'un logiciel (mais cela s'applique à la conception d'autre produit)

- ⇨ Expliquer la conception de l'architecture : le raisonnement et les décisions pourraient être tracés car ils sont liés aux objets de conception par une relation causale. (ce qui permettrait de ne pas omettre les contraintes et objectifs métiers, les critères environnementaux et éviter un traçage manuel des spécifications)
- ⇨ Evaluation de l'impact de la modification des exigences et des facteurs environnementaux : si une exigence est modifiée, la chaine des changements qu'elle va entrainer pourra être estimé par les traces.
- ⇨ Analyser des préoccupations transversales : les raisons des décisions proviennent de compromis multiples et parfois non reliés.
- ⇨ Tracer les causes premières : en cas de défaut : revenir aux exigences, aux hypothèses ou contraintes, aux choix et décisions
- ⇨ Vérifier l'architecture de conception a posteriori et garantir son intégrité: en l'absence des concepteurs, une vérification indépendante sera possible (et évitera une coûteuse reconstruction par rétro analyse)
- ⇨ Tracer l'évolution de la conception : quand des décisions sont prises, on a l'historique du raisonnement et le contexte des choix pourra être utilisé pour le futur

⇨ Mettre en relation des objets de conception : des objets disparates pourraient être réunis par une hypothèse, une exigence ou une contrainte commune (ce qui rejoint les préoccupations abordées dans la section 2.4(IR).

Il existe des méthodes de traçabilité qui incorpore partiellement la logique de conception. On peut citer les travaux de Haumer et al. (Haumer, 1999) qui suggère d'étendre le processus de conception afin de capturer et tracer les processus de décision à travers des artefacts comme la vidéo, les enregistrements audio, les schémas ou croquis. (Mais étant faiblement structurées par nature, ces information sont difficilement exploitables.) On peut également évoquer les travaux de Matta et al. (Matta, 2014) sur le recueil et la structuration du processus de prise de décision pendant les réunions.

La raison (ou le faisceau de raisons) sous-jacente à une décision ou un choix de conception est complexe à modéliser. La conception est le processus de synthèse parmi des alternatives dans l'espace de conceptions possible (Simon, 1981).

Si l'on reprend l'analyse de Yang (Yang, 2013) le raisonnement de conception (design reasoning) se manifeste sous 2 formes :

- Les raisons de motivation : ce qui motive l'acte de concevoir et donne un contexte (but à atteindre ou contrainte, par exemple une exigence, une hypothèse, ou un objet de conception, facteur environnemental)
- La logique de conception

La figure 2.2 (Yang, 2013) décrit un modèle conceptuel montrant les relations entre les motivations, la logique de conception, les objets de conception

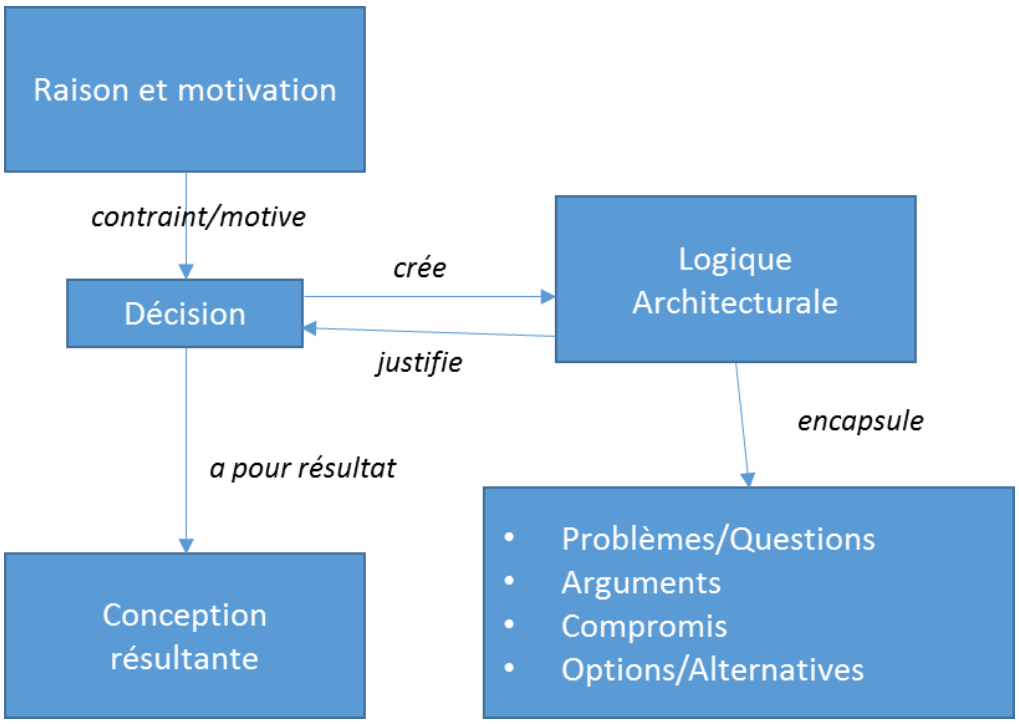


Figure 2.2 : d'après Yang, 2013

Ceci souligne bien l'importance de la traçabilité de la logique de conception en lien avec les objets de conception.

Enfin il nous parait intéressant de présenter quelques autres études visant à rapprocher la traçabilité et la logique de conception. On peut citer les ouvrages de B. Turban (Turban, 2013 et Turban, 2009) sur le sujet appliqué aux architecture logicielles qui reprennent certains des concepts précédents sur la traçabilité en y apportant une modélisation entité-relation tenant compte de la logique de conception.

Sur la figure 2.3, au lieu de simplement lier les exigences avec les contraintes pour en déduire des objets de conception (class), on rajoute un composant permettant la saisie des documents de décision

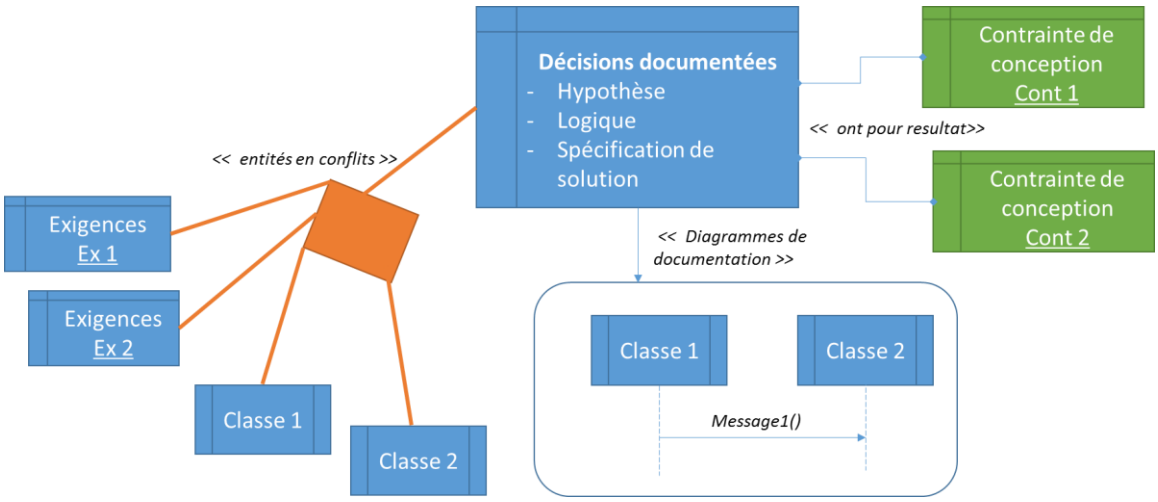


Figure 2.3 : Les décisions documentées établissent un pont entre les exigences, les éléments de conceptions et les contraintes de conception résultantes, d'après Turban 2009

Conklin (Conklin, 1989) affirme qu'en fournissant la logique de conception, la maintenabilité d'un système est accrue. Comme la traçabilité œuvre aussi dans ce sens Ramesh (1995) ont proposé un modèle de traçabilité et un outil (REMAP) (REpresentation and MAintenance of Process knowledge) qui combine IBIS avec les activité REM(Requirement Engineering Methodology)

Turban (Turban, 2013) rappelle que dans REMAP les objets traçables (exigences, éléments de conceptions, etc..) peuvent avoir 4 types de connexion vers le modèle décisionnel :

- La logique de conception est 'basée sur' les objets traçables
- Les hypothèses 'dépendent des' objets traçables
- Les décisions 'affectent' les objets traçables
- Les objets traçables 'génère' des problèmes ou des conflits

C'est en rapport avec les travaux de Knethen (Knethen,2001) et Pineihro (Piniehro, 2004) sur le CTM (Conceptual trace Model), un modèle entité-relation. Cette approche est intéressante par son apport à la modélisation des traces, le CTM consiste en deux types d'éléments principaux:

- Entité (type:produits intermédiaire, temporaires ou finaux; granularité; attribut (id, timestamp,etc..))
- Relation (quel type d'entité on trace, quel attribut (state, scope personne..) elles ont, avec quel granularité(niveau de détail),

La traçabilité s'effectuant principalement au niveau des relations qui ont des caractéristiques (Type, Direction (avant ou arrière), Attribut (auteur, date,etc..)). L'aspect original est que les types des relations peuvent être entre entité de même niveau d'abstraction ou à des niveaux d'abstraction différents (par exemple entre les exigences et la conception) et même entre différentes versions d'une entité. Nous retrouvons une des difficultés soulevées en partie 2.4 sur les techniques d'automatisation de la traçabilité.

Notre travail est dans la ligne de l'approche DYPKM (Bekti, 2003). Le processus DYPKM a pour objectif la collecte et la modélisation des connaissances visant à définir une mémoire d'expérience relative à un projet de conception dans un but de réutilisation. Pour ce faire un modèle relationnel global regroupant des éléments de contexte et de logique de conception est mis en place. Un mode de représentation de de modèle emploie la logique formelle et offre la possibilité de générer différentes vues sur la mémoire de projet. Les auteurs de DYPKM suggéraient qu'une piste de travail serait de s'intéresser à l'aspect communicationnel via la pragmatique (Bekti, 2003b pp. 116) et notre étude explore cette voie.

Il ressort de l'ensemble de ces travaux qu'il est crucial pour le cycle de vie complet du produit de lier traçabilité et logique de conception. Cependant avant de faire une synthèse de ces différentes approches, nous allons détailler une étape importante de la logique de conception : la résolution de problème.

2.7 La résolution de problème

La résolution de problème dans les projets industriels vise à transformer les connaissances de l'entreprise en valeur (Gray, 2001). Cela implique généralement deux types de connaissance: déclarative (sur des faits, des événements, et des objets) et procédurale (savoir comment faire les choses). Lors de la conception de logiciels complexes, les phases de résolution de problèmes se manifestent plus fréquemment, parce que les tâches sont décrites sommairement et souvent mal structurées par rapport à la conception d'un produit tangible.

Dans notre traçabilité de la logique de conception, nous avons d'abord choisi de nous concentrer sur la résolution de problème parce que les connaissances utiles pour l'entreprise sont plus susceptibles d'être utilisés ou créés au cours de ce type d'échange.

La théorie sur la résolution de problème a été développée à partir du travail de Newell et Simon (Newell, 1972) et a servi de base à beaucoup de travaux de recherche. Selon Hayes (1980) un problème surgit « chaque fois qu'il y a un écart entre l'endroit où vous êtes maintenant et où vous voulez être, et que vous ne savez pas comment trouver un moyen de combler cette lacune ». Par conséquent, trouver ce « chemin » est une partie importante de la résolution de problèmes. De manière analogue Metallidou (Metallidou, 2009) définit la résolution de problèmes comme un « comportement orienté vers un but qui exige une représentation mentale appropriée du problème et l'application ultérieure de certaines méthodes ou des stratégies afin de se déplacer d'une première, d'un état actuel à un état but désiré ».

Comme indiqué dans Mataka et al (Mataka, 2014), la psychologie cognitive a beaucoup apporté à la compréhension des processus mentaux en œuvre lorsque les individus apprennent et résolvent des problèmes. En particulier la nécessité d'une organisation de la connaissance afin d'améliorer la récupération de celle-ci depuis les schémas conceptuels lors de la résolution de problèmes (De Jong et Ferguson-Hessler, 1986).

La résolution efficace d'un problème commence d'abord par une compréhension de celui-ci via une description sous une forme facilement compréhensible pour aider la

recherche d'une solution appropriée (Reif, 1981). Cette description doit inclure des concepts clés nécessaires pour décrire les problèmes. Ensuite, selon les psychologues cognitifs, la résolution de problèmes englobe l'auto-analyse, l'observation, et le développement de procédés heuristiques (Hardin, 2002).

Selon Hardin (Hardin, 2002), " Tout problème a au moins trois composantes: les données initiales, l'objectif et les opérations ".

Cette définition générale de la théorie de la résolution de problème met en lumière les éléments clefs:

- Les « données » du problème: informations et les faits qui présentent le contexte;
- L'objectif: état final souhaité;
- Les opérations: les actions à réaliser pour atteindre l'état final;

Cela a conduit à l'élaboration d'approches cognitives à la résolution de problèmes. Hardin a développé un modèle par étapes de résolution de problèmes. Cela comprenait « (1) comprendre le problème, (2) concevoir un plan, (3) mettre en application le plan, et (4) regarder en arrière » (Hardin, 2002). Ces étapes génériques ne suivent pas un schéma linéaire et lorsque les solutions trouvées ne sont pas satisfaisantes, ce processus est cyclique.

Dans notre étude, nous allons nous concentrer davantage sur les données et objectif, à savoir la partie « de la reconnaissance du problème », les opérations faisant partie de la solution. Selon Bodner (1991) il existe deux catégories de problème : ceux que l'on a l'habitude de résoudre souvent, et ceux qui sont réellement nouveaux. C'est ce second type qui nous intéresse, et surtout les problèmes de type « indéfinis » ou « wicked problem » (Shum, 1997; Conklin et Weil, 1997), où la connaissance collaborative se manifeste naturellement.

Dans cette étude, nous cherchons à déterminer si la traçabilité de certains processus de communication en contexte et en lien avec les artefacts projets pourraient améliorer la compréhension de la résolution de problèmes en logique de conception.

2.8 Synthèse de ce chapitre

Le système d'information de suivi est une sorte de mémoire collective qui peut être utilisée pour accélérer la prise de décision dans les futurs projets. La traçabilité en lien avec la logique de conception a un rôle clef à jouer pour l'apprentissage organisationnel et la documentation des raisons derrière les choix et décisions critiques afin de transférer les connaissances vers l'entreprise.

Notre approche actuelle est basée sur une traçabilité a posteriori, c'est à dire une fois que le projet est terminé, et donc avec un risque que les membres de l'équipe projets ne soient plus disponibles.

Nous sommes davantage orientés vers une traçabilité arrière (depuis des produits ou services vers les artefacts de conception, les contraintes et les exigences).

Il est clair que dans notre cas des méthodes d'automatisation et d'IR pour générer des traces seront utiles car les équipes ne sont plus présentes et que de toute façon, il est difficile en environnement de production de rajouter des contraintes aux utilisateurs (à supposer qu'ils aient la capacité et le recul nécessaire) en leur demandant de définir eux même explicitement les liens entre artéfacts et exigences, et d'écrire l'historique de leur décision vis-à-vis des exigences. Nous allons nous placer dans une approche textuelle et regarder dans quelle mesure les communications médiatisées pourraient nous aider à reconstruire une partie de la traçabilité.

2.8.1 Notre Approche par rapport aux études existantes

Nous allons retenir des travaux précédents les résultats suivants :

- Les hypothèses, contraintes, raisonnements non documentés ou implicites ont besoin d'être redécouverts.
- Il est important de capturer la logique de conception de le lier avec la traçabilité des artefacts de conception pour l'évolution et la maintenance.
- Les méthodes actuelles incorporent la logique de conception de manière limitée et avec des informations peu structurées, il faudrait étendre cette liaison. L'étude des messages professionnels pendant un projet nous semble une piste prometteuse.
- TBR préconise une modélisation de la trace, nous allons essayer de modéliser une partie du contexte projets et des interactions au cours d'une phase de résolution de problème.
- De manière analogue à l'assistance utilisateur de TBR, nous allons en fonction d'un contexte de requête usager (d'une situation) rechercher dans les messages utilisateurs une trace de résolution de problème présentant des caractéristiques adéquates et pouvant contenir des connaissances
- Notre approche est résolument orientée Information Retrieval, le contexte global sera incorporé dans un système de calcul de score (voir chapitre 4 et 6) afin de trouver les situations similaires appropriées, ceci se rapprochant de la « signature » du TBR.
- A la différence du TBR et du CBR, nous n'allons pas faire de raisonnement, ni transformer les traces.
- Comme dans DYPKM, le contexte projet aura une grande part mais nous n'utiliserons pas la logique formelle
- Enfin à l'instar du TBR les traces seront partagées entre les utilisateurs ce qui facilitera le partage d'expérience.

Notre but est de retrouver d'utiliser des conversations médiatisées pour essayer d'y trouver des traces textuelles de logique de conception et de voir si on peut y trouver une structuration. Nous allons observer dans quelles mesure les emails professionnels peuvent être considérés à des fins de traçabilité, d'où la nécessité d'étudier les communications médiatisées et l'analyse du discours ce qui sera l'objet du chapitre suivant.

	TBR	CBR	DYPKM
Logique Formelle			*
Structuration		*	*
Signature cas	*		
raisonnement sur trace	*	*	
transformation	*	*	
contexte projet			*
communications projets	*		
logique conception			*
assistance utilisateur finaux	*		

Figure 2.4 : Quelques systèmes utilisant la traçabilité et en vert les aspects que nous allons retenir dans notre approche

Chapitre 3. PRAGMATIQUE DE LA COMMUNICATION MEDIATISEE ET EMAIL

3.1 La communication

Dans « Une logique de la communication » Watzlawick (1972) traite essentiellement de la pragmatique de la communication. Selon lui communication et comportement vont de pair : la pragmatique étudie les mots (la syntaxe), leurs sens (la sémantique), mais elle traite aussi des aspects non verbaux. Tout comportement est communication avec la fameuse formule « On ne peut pas ne pas communiquer » (indiquant que même une absence de comportement, de réponse, est une forme de communication).

3.2 La communication médiatisée

Dans le cadre de notre recherche, notre champ d'investigation sera celui de la communication médiatisée par ordinateur (CMO ou en anglais CMC Computer Mediated Communication) et plus spécifiquement de l'email ou courrier électronique. Historiquement la CMO s'est développée depuis une vingtaine d'année et elle a connu un très grand essor avec l'avènement de l'internet grand public, de la téléphonie mobile et du Cloud.

3.2.1 Définitions

La Communication médiatisée par ordinateur (CMO) pourrait être définie comme un processus dans lequel des êtres humains interagissent par transmission de messages par l'intermédiaire d'ordinateurs et de systèmes de télécommunication en réseau (Herring, 2002, 2004). En CMO l'échange entre les usagers se fait via le biais de logiciels et de différents types de technologies (protocoles) réseaux comme l'email (courriel), la messagerie instantanée, les forums, les réseaux sociaux (Facebook, Twitter), et la vidéoconférence (Skype) (plus anciens mais toujours actifs l'Internet Relay Chat (IRC), les mailings-lists (listes de diffusion) ou les serveurs de newsgroups).

Pour la majeure partie de la CMO, l'usage est basé sur texte (les messages sont tapés sur un clavier et affichés/lus comme texte sur un écran). Ainsi malgré des formes variées (email, chat en temps réels, groupe de diffusion) et des propriétés linguistiques propres à chaque type de de messagerie utilisé, l'activité qui se déroule via la CMO est constituée – dans de nombreux cas, exclusivement - par du langage représenté visuellement.

3.2.2 Caractéristiques

Par rapport à d'autres formes de communication, la CMO permet de s'affranchir des limites physiques et sociales et donc de permettre l'interaction entre des personnes ne partageant pas le même espace géographique.

Les principales caractéristiques de la CMO comprennent la persistance de la conversation (ou du moins sa capacité à être enregistrée), la communication plutôt formelle, et l'anonymat relatif des utilisateurs (n'importe qui peut être derrière un clavier ou un écran malgré les identités ou surnom affichés donc les notions d'identité sont plus floues). On peut également noter la nature multimodale de la CMO et le fait qu'elle ne soit pas spécifiquement adaptée à une gestion fine des codes de conduite sociaux ou culturels. En effet en l'absence de communications verbales (et/ou vidéo), l'interprétation des déclarations utilisateurs peut être difficile (Herring, 1994), (Walter, 1996).

On peut examiner et comparer la CMO avec d'autres médias en utilisant quelques caractéristiques (non limitatives) « universelles» à toute forme de communication comme la *synchronicité*, *l'anonymat*, et la *persistance* (McQuail, 1994).

La CMO existe en mode synchrone ou asynchrone. Dans la communication synchrone, tous les participants sont en ligne simultanément. Dans la communication asynchrone, il y a des contraintes de temps sur la prise en compte des messages et des réponses.

Selon le type de CMO, ces attributs seront associés différemment. Email et forums sont faiblement synchrones car le temps de réponse est très variable, mais ils offrent une persistance élevée car les messages envoyés et reçus sont sauvegardés. C'est également le cas des blogs où les commentaires entre l'auteur et les lecteurs sont affichés.

Au contraire la messagerie instantanée est intrinsèquement synchrone mais non persistante, car à moins d'activer un journal de log (trace ou de copier/coller celui-ci) le contenu de la conversation ne sera pas conservé à la fin du dialogue.

Enfin un autre domaine de recherche en CMO est celui de l´utilisation de caractéristiques paralinguistiques telles que les émoticônes, des règles pragmatiques telles que le tour de parole, l'analyse séquentielle et la structure des discussions. Ces aspects seront détaillés plus bas dans ce chapitre dans la partie sur la pragmatique.

3.2.3 Généralités

La communication écrite médiatisée par ordinateur (CMO) a été employée depuis de nombreuses années d'abord par les universités, les étudiants puis les entreprises et enfin les particuliers. Cependant la CMO de par ses particularités (problème de transmission des aspects non langagiers, distance, côté asynchrone, parfois anonymat (forums)) pose toujours la question de la relation interpersonnelle et en particulier des aspects sociaux affectifs.

La CMO a été étudiée par des linguistes (Baron, 1998) et des spécialistes de l'analyse du discours (Susan Herring, 1994) mais ce type d'approche privilégie l'aspect textuel et vise moins la composante émotionnelle de la relation. Les travaux de H. Atifi, N. Gauducheau et M. Marcoccia (Atifi, 2005) sur les forums de discussions permettent

de s'interroger sur les procédés sémio-discursifs utilisés pour la communication émotionnelle. Nous allons examiner quelques travaux choisis de S. Herring à travers lesquels nous avons abordé la CMO dans notre étude, puis nous examinerons ses usages professionnels et la place de l'email.

3.2.4 Travaux antérieurs

Susan Herring (1994,2001, 2002, 2004) a effectué un travail important sur la communication médiatisée par ordinateur. Dans une étude (Herring, 1994) elle présente quelques résultats sur la CMO, notamment la relative pauvreté du canal (pas audio/vidéo), l'anonymat, et surtout le peu d'investigation empirique des linguistes. Elle émet ensuite les hypothèses suivantes (basées sur des remarques des chercheurs comme Baron et Hale) et que l'on retrouve souvent quand la CMO est évoquée. La CMO à long terme:

1. réduirait la complexité syntaxique
2. réduirait le caractère soigné du message
3. réduirait la politesse
4. amènerait une variabilité réduite et davantage d'homogénéité syntaxique

Herring mène une étude linguistique quantitative sur un corpus basé sur un groupe de discussion ARPANET (1975-86) pour valider ces hypothèses. En utilisant des instruments de mesure linguistique (nombre de propositions subordonnées, de propositions complément, etc...) adaptés à chacune des hypothèses, elle calcule une variance de type ANOVA (Analysis of Variance) à 1 facteur. Les résultats montrent que les hypothèses 2, 3 semblent valides, mais pas les 1 et 4. (Notamment il n'y a pas d'augmentation du nombre de fautes, et ce résultat est encore plus valable depuis la généralisation des correcteurs orthographiques). Elle en conclut que les influences sociales sont autant à prendre en compte qu'un déterminisme technologique. Dans Herring (2004), elle indique que la CMO a besoin de méthodes provenant de l'analyse du discours. Ses méthodes sont adaptées de la linguistique et de la communication.

Ainsi l'étude de discours médiatisés par ordinateur est une spécialisation au sein de la plus large étude interdisciplinaire de CMO et se distingue par son utilisation des

méthodes d'analyse du discours. En 1994, Herring constatait que malgré le volume élevé d'activités online (forum, mail, blogs, chat, etc..) il existait paradoxalement peu de recherche basées sur de l'empirique, plutôt du spéculatif. Depuis de nombreuses études (Walther, 1996) (Schiffrin, 2008), (Herring, 2013) sont venues enrichir ce domaine.

Un des fondements de la CMDA (Computer Mediated Discourse Analysis) est que le discours présente des « patterns récurrents », et qu'il implique des choix du locuteur (qui reflètent des facteurs cognitifs et sociaux). La CMDA peut employer des techniques quantitatives ou qualitatives. Elle doit cependant poser une question de recherche à laquelle on puisse répondre, motivée par une hypothèse, proposer des méthodes de comptage, des catégories, un corpus valide. Enfin les concepts clef doivent être « opérationnalisables » c'est à dire qu'un chercheur avec les mêmes données doit pouvoir identifier et quantifier un concept sans ambiguïté. On parle ici de phénomènes textuels que l'on peut observer, coder, compter (en clair le plus un concept est abstrait, interne, le moins il sera « opérationnalisable »). Selon Herring (2002), la CMDA s'applique à 4 domaines/niveau du langage : structure (mot, orthographe..), sens (sémantique, acte de langage..), interaction (tour de parole,..), comportement social (jeux, pouvoir, conflits..). Ses méthodes sont adaptées de la linguistique et de la communication. Herring présente une approche basée sur « coding and counting ». Ce dernier travail a le mérite de mettre en place un cadre clair et une méthodologie précise tout en rappelant que le codage de certains phénomènes ne pourra être fait que de manière manuelle.

3.2.5 Usages de la CMO en entreprise

La CMO est réputée plus impersonnelle de par son origine (elle a été conçue au départ pour relier des ordinateurs mainframe pour des raisons de sécurité et de redondance d'information) et le fait de pouvoir envoyer et recevoir des messages à distance par écran interposé n'était qu'un effet de bord. Par la suite, elle a été employée pour la communication et le travail de de groupe. Rapidement la question s'est posée de son efficacité par exemple par rapport aux réunions et méthodes traditionnelles (Turroff, 1991). Malgré cette réputation, la CMO n'est pas forcément impersonnelle et peut

parfois revêtir un caractère aussi personnel qu'une conversation en face à face (Face to Face ou FTF en anglais).

La CMO est particulièrement adaptée à la coordination et l'usage de l'email et de la vidéoconférence ont réduit les problèmes d'affects. Walter (1994) a même montré que la CMO était plus orientée « tâche » que les réunions FTF. C'est un aspect positif du caractère impersonnel car il amène une forme de rationalité et de discipline particulièrement adaptée à la gestion et à la réalisation des tâches. En particulier, la CMO en entreprise apporterait selon Dubrovsky (1995):

- Une meilleure coordination, une participation des intervenants plus équilibrée, une meilleure gestion et réalisation des tâches, un « filtrage naturel » des aspects affectifs, une meilleure mise en commun des contenus et une minimisation des influences sociales (statut, rôle).

Dans la cadre qui nous intéresse de la résolution de problème de manière collaborative en gestion de projet, la CMO peut aider et améliorer la décision de groupe. En effet, le caractère impersonnel que l'on reproche souvent à la CMO provient essentiellement du manque de non verbal. Paradoxalement dans un cadre professionnel cela peut se révéler parfois bénéfique : Kiessleer (1986) a montré que sans le non verbal un locuteur peut difficilement altérer le ton du message, exercer de la dominance ou du charisme. Mais cependant des stratégies linguistiques visent à pallier à ce manque comme nous le verrons dans la partie pragmatique. Cela dépend du potentiel des utilisateurs à s'adapter à un code linguistique. En revanche, si l'on se limite à un critère temporel pur en temps limité la CMO mettrait plus de temps à atteindre un consensus que le FTF.

Par ailleurs, Watzlawick (1972) définit également deux notions très importantes : la *symétrie (*les partenaires ont des comportements en miroir, avec égalité et minimisation de la différence) et la *complémentarité* (le comportement de l'un des partenaires complète celui de l'autre, avec maximalisation de la différence). En effet en termes de communication, il est légitime de se demander comment cela se passe-t-il pour les locuteurs, à savoir l'émetteur et le récepteur. (Nous nous limitons dans cette section à un seul mais l'email peut avoir simultanément plusieurs dizaines de destinataires)

En l'absence d'indices non verbaux et/ou de rencontre physiques l'émetteur va optimiser sa présentation et avoir une perception que l'on pourrait qualifier d'idéalisée du récepteur. Plus précisément comme l'indique Walther (1996) :

- L'émetteur peut sélectionner et exprimer des comportements qui sont plus désirables ou appropriés (selon lui) pour atteindre son but et transmettre un message libre de « bruit »
- Au niveau du récepteur, il construit une image idéalisée de son interlocuteur et de leur relation et ils se les confirment réciproquement. Le caractère asynchrone facilite encore cela car il peut supprimer en partie la pression et le stress du temps réel. Le temps est gelé entre les interventions. Ce point de vue est à relativiser suivant les travaux d'Atifi (2011) pour qui les emails peuvent au contraire accroitre la pression temporelle, cette dernière approche semble se confirmer avec les propositions sur le droit à la déconnexion dans le projet de la nouvelle Loi travail[3].

3.2.6 CMO et traçabilité de la résolution de problème

Pour en revenir au sujet de notre étude, la CMO apporte énormément à la collaboration par ordinateur, et particulièrement à la résolution de problème de groupe (Jonassen, 2001). En effet l'efficacité de celle-ci est largement déterminée par le niveau de communication des membres du groupe. Ces membres interagissent pour s'influencer, la communication est ce qui influe le plus sur l'efficacité des décisions dans le cheminement vers la résolution (Hirokwa et Pace, 1983). Bien que le manque du non verbal rende moins personnel et diminue le côté social/affect, cela facilite la gestion et l'exécution des tâches (Walther, 1996)

On peut se questionner sur la manière dont la CMO agit sur la structuration de la communication en situation de « problem solving » comparée au FTF.

Nous avons vu dans le chapitre précédent que le « problem solving » dans les entreprises contient souvent des problèmes mal structurés (des buts peu clairs ou mal

[3] http://travail-emploi.gouv.fr/grands-dossiers/projet-de-loi-travail/quelles-sont-les-principales-mesures-du-projet-de-loi-travail/article/droit-a-la-deconnexion

définis, des contraintes pas toujours établies, et des solutions multiples (voire aucune)), donc des « wicked problems » (au sens Buckingham- Schum (1997), Rittel et Webber (1973)).

La CMO est d'un grand apport dans ce contexte-là : elle amène une meilleure prise de conscience de la situation globale, une autorité davantage basée sur la connaissance que sur la hiérarchie organisationnelle et surtout un processus de design créatif beaucoup plus abstrait (en raison de la distance, du caractère asynchrone, davantage de réflexion et de prise en compte des apports des autres, des demandes de tous, des référentiels communs, et le développement de communautés d'intérêt. On peut récapituler les changements majeurs apportés par la CMO dans ce cadre :

Elle apporte comme bénéfices (Zellhofer 1998), (Kiesler, 1984) (Lane, 1994) :

- Une disparition des barrières géographiques.
- Un échange, un stockage, et des transferts de document instantané.
- Peu d'indices sur les statuts/position, respect hiérarchie, et donc davantage d'égalité et une « hiérarchie de la connaissance ».
- Elle peut accroitre la taille efficace maximum dans les groupes de travail (travailler à 50 sur un projet).
- Paradoxalement, l'utilisation du groupware conduit à de nouvelles façons d'aborder les réunions en F2F.
- Les communications rapides réduisent les délais, les organisations (et les personnes) apprennent davantage et plus vite à propos des événements qui les intéressent.
- Davantage d'égalité dans la participation que dans les médias conventionnels.
- La CMO promeut l'égalité et la flexibilité des rôles de chacun.
- Elle accroit les opportunités de communication décentralisée, le contenu des fils de conversation augmente.
- La CMO peut augmenter la communication informelle, cela accroit la connectivité du réseau de communication.

Mais avec quelques effets de bord (Zellhofer 1998) (Lane, 1994):

- Elle véhicule mal ou pas du tout le non verbal, social, culturel

- Changement de structure de communication officieuse dans les organisations : elle accroît la création de réseaux latéraux à l'intérieur d'une organisation ou entre organisation. Ce faisant elle modifie les structures sociales en passant d'un système pyramidal ou hiérarchique à un modèle en réseaux. En termes de SNA (Social Network Analysis ou analyse des réseaux sociaux) cela a un impact sur la centralité des membres dans un groupe.
- La régularité des interventions/participations de certains membres est parfois plus difficile à obtenir
- En terme de leadership/management : La CMO augmente le besoin d'un leadership fort et actif mais l'émergence d'un leader est moins probable et se fait différemment.
- Elle augmente le risque potentiel d'une élite numérique, et accroit l'étendue des responsabilités.
- Les questions restent parfois sans réponse (ce qui n'arrive pas en FTF)
- Les groupes parlent plus longtemps pour atteindre un consensus et/ou trouver des solutions
- La focalisation sur un thème est parfois plus difficile que dans les discussions FTF

3.2.7 CMO et Email

N. Baron (1998) explique bien la différence entre voir quelqu'un et voir l'image de quelqu'un à travers des messages. Ce type de questionnement est rejoint par Panckhurst (1998,1999) lorsqu'elle indique que l'email contient davantage d'erreurs (typographiques et/ou orthographiques), moins de relecture, un discours plus « relâché », et un style différent du courrier épistolaire.

Ce point de vue est réfuté par Marcoccia (2003) qui indique dans son étude des genres et typologie en CMO, le message professionnel a ses codes aussi (peut apparaitre sous forme de memo, proposition, ballot (vote), dialogue). Il indique que certains dispositifs comme le forum appartiennent à un genre hybride (à la fois communication de masse et interpersonnelle) mais l'email hérite plutôt du genre de la lettre. Afin de

mieux cerner la place de l'email au sein de la CMO comme nous allons l'examiner plus en détail dans la partie suivante.

3.3 L'Email

3.3.1 Définition et historique

L'histoire de l'email ou « courrier électronique » (ou encore courriel, note de bas de page) date de 1970 (et d'ARPANET) où il était employé dans les cercles gouvernementaux, militaires ou académiques. Il permet d'envoyer un message texte via un réseau d'ordinateur. Il crée une forme de communication différente, puisque l'on peut envoyer un email à une personne que l'on n'aurait pas forcement contacté par téléphone ou lettre.

Dans Tao (1996), on retrouve la définition officielle « Email is the short form for electronic mail. By definition it is mail delivered through electronic means ». Avec l'évolution des moyens technologiques l'email actuel transporte non seulement du texte mais également des fichiers, images ou vidéo et il est accessible via des interfaces web et/ou des téléphones cellulaires.

3.3.2 Qu'est-ce qu'un e-mail ?

Un courrier électronique (ou e-mail) est un petit paquet de données qui transite sur Internet, d'un ordinateur à un autre, plus précisément d'une boîte aux lettres électronique à une autre. Il est composé :

- de l'**adresse** électronique du ou des destinataires principaux (et éventuellement des destinataires en copie)
- d'un **contenu** (dit aussi *corps*), qui contient par exemple le texte du message, des fichiers (comme des images), etc.
- de quelques **informations** destinées à l'acheminement du message, équivalentes des tampons postaux (date et heure, etc.).

A la différence du courrier postal, le courrier électronique comporte en plus un **objet** (ou sujet) pour identifier le message.

Source http://www.arobase.org/bases/introduction.htm, consulté en juillet 2016.

Formellement un email est un type message comprenant :

- Des données : un sujet, un corps de texte (ou HTML) et éventuellement des pièces jointes.
- Des métadonnées : un émetteur, un ou des récepteurs (éventuellement en copie ou caché), une date d'envoi et de réception. (et dans son en tête (Header) des informations permettant de tracer son trajet)
- En outre il s'inscrit souvent dans un fil (thread) comprenant des messages et les réponses associées.
- Il utilise des protocoles spécifiques (SMTP, MIME, POP, IMAP, etc....)

Un système d'email utilise des ordinateurs afin de créer un environnement d'échange d'information à haute vitesse (intra ou inter site). L'utilisateur utilise un logiciel client email (comme Outlook, Thunderbird, etc..) pour écrire ou lire ses messages qui seront relayés via des serveurs au(x) destinataire(s). Pour résumer l'email est asynchrone, rapide, textuel.

3.3.3 Statistiques et Chiffres

Selon le site arobase.org et le groupe Radicati[4] (Email Statistics Report, 2015-2019) en 2015, on atteint un total de 2.6 milliards d'utilisateurs pour 205 milliards d'email quotidiens (hors spam, (email non sollicité ou « pourriel »)), c'est devenu un moyen de communication aussi courant que le téléphone. L'email vient en deuxième position des services Internet après la consultation de sites. Par exemple dans les formulaires en ligne, ce que l'on demande c'est une adresse email et il est tenu pour acquis d'en avoir au moins une. Comparé à d'autre mode de CMO, l'email est un des plus anciens (en perte de popularité au profit des messageries instantanée) mais reste prédominant, surtout en entreprise.

[4] http://www.radicati.com/?p=12960

88 courriels sont reçus et 34 sont envoyés en moyenne par jour en entreprise par chaque collaborateur. Les courriers électroniques augmentent sensiblement le volume des communications dans l'entreprise (malgré les nombreux moyens de filtrage de spam mis en place) (arobase.org[5], 2016).

En France le téléphone arrive encore en tête avec 41 communications par jour, mais l'email est en augmentation constante et certains managers consacrent parfois plusieurs heures par jours à traiter leur boite d'emails. Une enquête[6] de 2007 par le groupe Datamonitor, réalisée auprès de 400 responsables informatiques employés dans 524 entreprises de 13 pays différents, montrait déjà qu'ils étaient désormais 100% à utiliser le courrier électronique, contre 80% pour le téléphone fixe, 76% pour le téléphone portable, et 66% pour la messagerie instantanée.

Daily Email Traffic	2015	2016	2017	2018	2019
Total Worldwide Emails Sent/Received Per Day (B)	205.6	215.3	225.3	235.6	246.5
% Growth		*5%*	*5%*	*5%*	*5%*
Business Emails Sent/Received Per Day (B)	112.5	116.4	120.4	124.5	128.8
% Growth		*3%*	*3%*	*3%*	*3%*
Consumer Emails Sent/Received Per Day (B)	93.1	98.9	104.9	111.1	117.7
% Growth		*6%*	*6%*	*6%*	*6%*

Figure 3.1 : Traffic Email Mondial en milliards (B) d'après Radicati Email Statistics Report, 2015-2019 – Executive Summary

L'abondance des messages peut distraire et interrompre fréquemment les employés et nuire ainsi à leur travail. En outre de plus en plus de salariés consultent leurs emails professionnels en dehors des heures de bureau. (Xobni, septembre 2010)

Tout en étant moins personnel qu'une lettre manuscrite, c'est cependant un moyen de communication qui interpelle moins directement (par exemple comparé au téléphone)

[5] http://www.arobase.org/actu/chiffres-email.htm

[6] http://www.lemondeinformatique.fr/actualites/lire-l-email-premier-outil-de-communication-en-entreprise-23739.html

mais qui reste relativement intrusif. Pour les personnes très éloignées géographiquement, il garantit une livraison sûre et instantanée des messages. Il permet de délivrer la même information à plusieurs personnes (par rapport au téléphone), de pouvoir prendre le temps de réfléchir à sa formulation et sa réponse. Pour des personnes peu adeptes de l'écrit papier ou de la communication FTF, sa facilité d'usage est incitative. Il est singulier de constater que des employés qui travaillent dans le même bureau à quelques mètres de distance communiquent essentiellement par email. C'est, en entreprise, un formidable accélérateur de diffusion d'information (qui sinon aurait été reçue plus lentement voire pas du tout) (Crawford, 1982). Dans un contexte professionnel marqué par la fragmentation et la multi activité (Licoppe, 2008), l'étude d'Atifi (2015) indique que les utilisateurs peuvent déployer des stratégies linguistiques afin de donner une force d'interpellation importante à leurs messages électroniques et faciliter ainsi la prise en compte par les destinataires

3.3.4 Email : quel genre ?

Avant l'ère moderne, les communications étaient assez simple : soit directes (face à face) soit distante (courrier), mais avec l'évolution des moyens technologiques et le numérique, on utilise le fax, le téléphone mobile, et surtout l'email (en particulier dans le monde professionnel). Et les linguistes se sont alors posé la question : l'email est-il une nouvelle variété de langage parlé ? Écrit ? Ou encore d'une autre nature ?

L'email étant au centre de notre travail, il convient de se questionner sur sa proximité avec les genres écrits (lettres) et oraux (conversations) afin de choisir les modes d'analyses et les méthodologies les plus appropriés.

L'email suit des conventions de la conversation orale mais également celles de la lettre. Malgré son origine épistolaire et sa forme écrite, il entre dans une structure conversationnelle de par la rapidité de rédaction et de transmission des messages.

On retrouve des similitudes avec les travaux sur les forums de discussion : Collot et Belmore (1996) ont examiné des messages de forums. Ils ont trouvé qu'ils ressemblaient plutôt à des entretiens ou à des lettres, et dans la composante interactionnelle et personnelle se rapprochaient plus du genre parlé que de l'écrit.

Si l'email est un genre distinctif qui résulte de l'apparition d'un nouveau moyen de communication, les chercheurs tentent cependant de le rapprocher de genres existants de par sa forme et son contenu. Nous projetons d'utiliser des techniques d'analyses conversationnelles sur les emails aussi nous allons valider cette approche en examinant les travaux à ce sujet.

Naomi Baron (1998) évoque des caractéristiques linguistiques du courrier électronique assez semblables à la parole: style informel, simple et souvent ingénu, et un sentiment pour l'usager d'un media éphémère. Elle décrit d'abord les différences entre langage parlé et écrit notamment séparation spatiale/face à face, monologue/dialogue, durable/éphémère et formel/informel. Elle en conclu que l'email emprunterait un peu aux deux (Baron, 1998). En 2003, elle explore les façons dont la technologie a modifié notre façon d'écrire : le contenu (le fond) prenant le pas sur la ponctuation, l'orthographe. La ligne écrit/parlé s'estompant peu à peu au niveau du genre, l'email se rapproche du langage parlé (Baron 2003).

Herring (1996) indique aussi que l'email est du type « sens unique » asynchrone, contrairement aux dialogues usuels en FTF, ou encore le téléphone et les messageries instantanées qui sont à double sens et synchrone. Aussi, bien qu'il emprunte beaucoup au parlé de par sa forme, les conversations et les interactions se font en aveugle et avec des morceaux de temps figé. Le langage oral est celui de l'immédiateté et l'écrit celui de la distance (Koch et Oesterreicher, 1994). Le rapport à l'espace, au temps, et à l'autre intervient dans le choix du média de communication : en cas de distance spatiale ou de distance sociale (nécessité d'un aspect formel), l'écrit sera plus approprié. Herring mentionne la notion de « conceptuellement parlé », car en pratique un email est bien toujours de l'écrit, mais c'est *conceptuellement* que la ligne écrit/parlé s'amoindrit. L'email professionnel est moins conceptuellement parlé qu'un email privé à un ami. Les choix de communication ne sont pas fixés et s'adaptent. Les modalités d'utilisation du langage contribuent également à la création d'un contexte, c'est donc échange (Verschueren 1999).

Crystal (Crystal, 2001) a introduit le terme de « Netspeak » pour désigner l'usage de la langue dans la CMO. Il constate que l'email est un genre encore en évolution et qu'il n'y a pas vraiment de langage spécifique au courrier électronique mais plutôt de

nouvelles pratiques, par exemple les dialogues par emails sont spécifiques via les fils (threads), avec la reprise des messages précédents et la possibilité de répondre à l'intérieur d'un message pour recréer un forme de dialogue a posteriori (simulation de l'interdialogue). Un quasi dialogue est réalisé dans ses fils et justifie des outils d'analyse du discours. Pour lui « Netspeak doit être vu davantage comme un langage écrit qui a été tiré vers le parlé que comme un langage parlé mis par écrit ». Une certaine étiquette non écrite prévaut cependant dans les échanges : par exemple ne pas retransmettre (« forwarder ») un message sans l'accord de l'émetteur initial, éviter de révéler les adresses de tous les destinataires dans un envoi très large, éviter aussi de répondre à tout le monde dans un tel cas. Mais cela relève davantage du savoir-vivre numérique que d'un comportement codifié.

Panckhurst (1998, 1999) étudie 1285 courriers électroniques universitaires entre 1997-1998 afin de montrer dans quelle mesure l'outil médiateur façonne le discours électronique. Ce travail vise à établir que le mail induit une nouvelle forme de discours. Les méthodes employées comportent des comptages de marques, pronom, classement des verbes, temps. Les résultats indiquent que le courrier électronique apporte un discours plus souple, informel, différent du style épistolaire, avec parfois davantage d'erreur typographique ou d'orthographe (mais ceci n'est plus valide avec la généralisation des correcteurs orthographiques), le « ne » de la négation reste et on note là aussi une simulation de l'interdialogue en écrivant sa réponse au milieu du mail reçu. Panckhurst pose la question de savoir si les intentions illocutoires passent bien via le medium du courrier électronique car il semblerait que des distorsions soient fréquentes.

De même, l'email de par sa nature écrite et distante oblige les utilisateurs à mettre en place des procédés de représentation du non-verbal (Yates, 1993), (Marcoccia 2004)

Marcoccia (2004) indique : « Faire du face à face avec de l'écrit pose problème puisqu'une partie du matériau sémiotique disponible dans la conversation en face à face disparaît naturellement avec la communication médiatisée par ordinateur (..) poster un courrier électronique comme si c'était une conversation en face à face revient bien souvent à imaginer un certain nombre de procédés permettant de représenter le non-verbal et le para-verbal ».

3.3.5 Usage de l'email en milieu professionnel

Bien qu'il soit traditionnellement dédié à une communication plus sure et efficace, on peut essayer de recenser quelques usages pratiques de l'email dans un groupe, au travail, dans un projet. Oren Ziv (1996) a étudié comment des personnels universitaires vont employer l'email pour négocier les conflits techniques et organisationnels afin d'accomplir des tâches. En particulier il aborde la façon dont les différentes formes de communication interagissent (email, téléphone, réunion, FTF, etc..). Certaines formes de communication sont plus adaptées au respect de la hiérarchie officielle. On peut voir dans la table reproduite figure 3.2 les différents usages de l'email dans un projet de groupe en milieu professionnel.

De cette étude on peut retenir que le courriel en milieu professionnel académique sert surtout à planifier d'autres rendez-vous (réunion, téléphone, face to face, etc..), et qu'il n'aplatit pas la hiérarchie mais la reflète, permet son extension, et également qu'il a un rôle social (d'appartenance à la structure).

Communicative Purposes	*E-Mail*	*Printed Documents*	*Team Meetings*	*Direct Meetings*
Acquiring Technological Expertise			=	
Assigning New Responsibilities				+
Collaborating and Cooperating			+	
Considering Cross-training			=	+
Considering Information Storage/Access			=	
Considering Site Location			=	
Delivering Documents	-	=	=	
Discussing Goals and Objectives		=	+	=
Discussing Hiring Process			+	
Discussing Information Sharing			+	
Discussing Using E-Mail	-		=	=
Establishing Links with AIS			=	+
Forwarding Information	+			
Handling Administrative Requests	=	=		
Improving Health and Safety			=	
Offering Feedback	=		-	+
Personal Correspondence	=			
Preparing Budget		=	+	+
Prioritizing Responsibilities			-	+
Providing Opinions	-		-	+
Reporting Information	+	=	+	+
Reporting Project Status	=		+	+
Requesting Action	-	=	-	+
Requesting Information	+		=	+
Reviewing Documents	-		-	+
Setting Schedules	-		-	+
Solving Computer Problems	+		=	+
Suggesting Meeting or Phone Call	+		-	

Key
Often + Seldom -
Occasionally = Not Observed

Figure 3.2 : Tableau d'O. Ziv sur une taxonomie des objectifs selon les moyens de communication dans son étude

3.3.6 L'email dans notre étude

En accord avec les travaux de S. Herring évoqués plus haut et ceux de Cohen (2004), nous croyons que l'email se rapproche assez de la conversation ou se situe dans un cadre interactionnel pour chercher à l'analyser en termes d'actes de discours. Toutefois notre approche diffère de celle de Cohen car nous concentrerons sur les actes et leur combinaison avec les caractéristiques spécifiques de l'email (métadonnées) et le contexte global de l'échange (dans notre cas le projet).

La CMDA décrite par Herring (2001) permet d'analyser le contexte dans lequel ces interactions apparaissant. L'importance du contexte, des intervenants, de la relation qu'ils établissent est primordiale dans cette perspective. Dans ce cadre, la pragmatique de la communication nous est apparue comme un bon révélateur. La pragmatique est une approche fonctionnelle du langage. Elle suppose l'étude du langage en envisageant son utilisation pour la génération de sens. Cela implique que tous les niveaux du langage et les phénomènes linguistiques associés sont analysés en prenant en compte :

- les traits/caractéristiques approprié(e)s du contexte dans lequel ils/elles occurrent
- la façon dont ces facteurs contribuent à fabriquer (faciliter la co-construction) du sens entre les locuteurs.

Nous allons décrire plus précisément l'apport de la pragmatique et de l'analyse du discours à notre étude dans les parties suivantes.

3.4 La Pragmatique et l'email

La théorie des actes de langage découle des œuvres originales en philosophie du langage d'Austin (1962) et Searle (1969). Depuis de nombreuses études ont été menées sur les actes de langage, et en particulier la requête, dans différentes disciplines comme la linguistique théorique et le traitement du langage naturel (NLP Natural langage processing). Par exemple, on peut citer les travaux de Cohen (2004) et Carvalho (2006) sur l'automatisation de l'identification des actes de langage dans des courriels. Rachele De Felice (2012,2013) a même proposé un schéma de classement détaillé pour annoter les actes de langage dans un corpus d'emails professionnels mais a cependant noté qu'une majorité de travaux sur les actes de langage portent sur le langage parlé. La pragmalinguistique étudie la façon dont les formes linguistiques sont utilisées dans l'exécution de certains actes de langage.

3.4.1 Acte de Langages (AL)

Le philosophe John Austin a posé comme hypothèse que le langage n'a pas qu'une fonction descriptive mais sert aussi à accomplir des actes. Dans les énoncés :

1. La porte est ouverte.
2. Je te promets venir travailler sur ce projet.

Austin qualifie le premier énoncé de constatif (décrit le monde), alors que le second est performatif (accomplit une action). Ainsi un locuteur accomplit un acte (dit Acte de langage, **AL** ou Speech Act) simplement en prononçant un énoncé (Austin, 1962). En CMO textuelle, la « prononciation » est le fait de taper sur le clavier un acte (qui est exécuté dans le monde réel ou le monde virtuel en co-construction) par exemple « merci ».

3.4.2 Performatifs, actes locutoire, illocutoires, perlocutoires

L'objectif de la pragmatique est d'analyser la manière dont les locuteurs utilisent la langue pour effectuer des AL, par exemple, des demandes, des excuses, des commandes, des promesses, des conseils, etc.

Les performatifs sont au cœur de la pragmatique. Ils se manifestent quand les locuteurs exécutent des actes simplement en les énonçant (« Quand dire c'est faire » (Austin 1962)). Les performatifs consistent en des actes locutoires (les mots prononcés) et illocutoires (la force de ces mots dans la situation de communication). Plus précisément Austin distingue trois niveaux de signification des actes de langage:

- Locution - au sens littéral
- Illocution - le sens voulu par le locuteur
- Perlocution - l'effet de l'AL sur le destinataire

En effet tout orateur qui fait un énoncé significatif (qu'il soit ou non performatif) tente d'effectuer un acte illocutoire dans le contexte de son énoncé. Si ces auditeurs réagissent (par exemple en étant convaincus, en obéissant, ou en refusant) ils accomplissent ainsi des actes perlocutoires. C'est dans ces tentatives de performance des actes illocutoires que les locuteurs expriment et communiquent leur pensées dans la conduite du discours.

3.4.3 Typologie des AL et Force Illocutoire

Austin avait suggéré cinq types d'actes illocutoires et identifié des verbes associés. Son approche a été complétée par d'autres auteurs (Voir le tableau figure 3.3 pour une comparaison).

Table 1: Comparison of Traditional Speech Act Categories by Author.

SA Category	Austin	Vendler	Bach & Harnish	Description
Assertive	Expositive	Expositive	Constative	expound views, state, contend, insist, deny, remind, guess
Commissive	Commissive	Commissive	Commissive	commit the speaker: promise, guarantee, refuse, decline
Behabitive	Behabitive	Behabitive	Interpersonal	reaction to others: thank, congratulate, criticize
Interrogative		Interrogative	Directive/Query	ask, question
Exercitive	Exercitive	Exercitive	Directive/Request	exercise power, rights or influences: order, request, beg,
Verdictive	Verdictive	Verdictive & Operative		giving a verdict: rank, grade, define, call, analyze

Figure 3.3 : Tableau des catégories d'AL

Selon l'approche logique standard, un locuteur utilisera des propositions avec des conditions de vérité alors qu'en philosophie du langage ce seront des actes illocutoires avec des conditions de réussite (félicité). Ces actes s'accomplissent par l'énonciation d'une phrase appropriée dans un contexte adéquat.

Ce court aperçu de la pragmatique ne serait pas complet sans mentionner les contributions faites par Searle (1969). Ce dernier a révisé la classification d'Austin en introduisant les termes de « contenu propositionnel» (le sens littéral d'un énoncé) et «force illocutoire » (ce que le locuteur entend par ce qui est dit).

Un locuteur fait un énoncé en prononçant ou écrivant une phrase. En tentant de réaliser un AL, il exprimera sa proposition avec une force illocutoire. Ainsi selon cette approche les AL sont de la forme F(P): ils sont composés d'une force F et d'une proposition P. Par exemple :

Les phrases « viendrez-vous demain ? » et « est-ce qu'il pleut dehors » ont la même force illocutoire F mais des contenus propositionnels différents.

De manière analogue des phrases « s'il vous plait ouvrez moi la porte » et « vous ouvrirez cette porte » ont des contenus propositionnels identiques mais des forces illocutoires différentes.

3.4.4 AL et Email

La théorie des AL associe aux unités de discours (« utterance » qui n'est pas forcément une phrase) un ou plusieurs actes de langages explicites. Un courrier électronique peut être considéré comme une séquence d'un ou plusieurs énoncés textuels et ainsi comme une séquence d'actes de langage. Dans un cadre professionnel, en caractérisant un email par son AL le plus important, cela pourrait permettre de le catégoriser en fonction de l'action espérée par l'expéditeur vis-à-vis du destinataire.

Cette démarche a été appliquée au courrier électronique dans différents travaux, par exemple :

- Gestion des tâches et des engagements par email (Kalia (2013), Lampert (2007) Khoussainov (2005))
- Classification selon intention de l'expéditeur (Cohen(2004), Carvhalo (2006), Goldstein (2006))
- Détection des sujets de discussion dans les fils de conversation (Feng (2006))
- Découverte des rôles des utilisateurs dans l'organisation (Leuski(2004))
- Analyse de l'efficacité des emails professionnels (Atifi et al. (2011))
- Taggage /zonage de segments de discours dans les emails (Yelati (2011), Lampert (2010))

Nous reviendrons sur certains aspects techniques ces travaux dans le chapitre suivant consacré au traitement du langage naturel et en particulier à l'email.

3.4.5 Actes de langage directs et indirects

Dans «Indirect Speech Acts», Searle(1975) fait la distinction entre les actes de langage directs et indirects. Les AL directs sont ceux dans lesquels le contenu propositionnel porte la force illocutoire.

Il définit un acte de langage indirect comme « un acte effectué par l'intermédiaire d'une autre » et déclare que, dans le discours indirect le locuteur communique davantage que ce qui est réellement énoncé. Ainsi, dans les actes de langage directs, il existe un lien entre le sens littéral et le sens classique, ou entre la forme et la fonction

de l'énoncé. Dans les actes de langage indirects, le sens littéral et le sens classique sont différents.

Afin d'expliquer au mieux comment le locuteur peut générer des AL indirects et comment son auditeur arrivent à les interpréter, Searle suggère qu'ils partagent les mêmes informations de fond linguistiques et non linguistiques leur permettant de faire des inférences correctes.

La force illocutoire de certains actes de langage indirects peut être interprétée en fonction de leur utilisation par exemple:

- la capacité du destinataire à accomplir un acte
- le souhait/besoin de l'émetteur
- la suggestion concernant l'auditeur
- le désir ou volonté du destinataire à faire un acte

On peut regarder le tableau de la figure 3.4 issu des travaux de Trosborg(1995) ci-dessous sur les stratégies de requêtes indirectes sur lesquelles nous reviendrons plus loin.

Table 4.1 Request strategies (presented at levels of increasing directness)

Situation: Speaker requests to borrow Hearer's car.

Cat. I	**Indirect request**		
	Str. 1 Hints	(mild)	I have to be at the airport in half an hour.
		(strong)	My car has broken down.
			Will you be using your car tonight?
Cat. II	**Conventionally indirect**		
	(hearer-orientated conditions)		
	Str. 2 Ability		Could you lend me your car?
	Willingness		Would you lend me your car?
	Permission		May I borrow your car?
	Str 3 Suggestory formulas		How about lending me your car?
Cat. III	**Conventionally indirect**		
	(speaker-based conditions)		
	Str. 4 Wishes		I would like to borrow your car.
	Str. 5 Desires/needs		I want/need to borrow your car.
Cat. IV	**Direct requests**		
	Str. 6 Obligation		You must/have to lend me your car.
	Str. 7 Performatives		
	(hedged)		I would like to ask you to lend me your car.
	(unhedged)		I ask/require you to lend me your car.
	Str. 8 Imperatives		Lend me your car.
	Elliptical phrases		Your car (please).

Figure 3.4 : Stratégie de requêtes en fonction de leur caractère direct croissant d'après Trosborg (1995)

Ces structures indirectes nous intéressent particulièrement dans notre étude car elles se produisent fréquemment dans le milieu professionnel et dans les emails par le biais du respect de la hiérarchie et de la politesse linguistique et que nous allons détailler ci-dessous.

3.4.6 Les AL et la politesse

Une des raisons pour lesquelles il y a de multiples façons d'accomplir des actes de langage indirects est la politesse (Brown et Levinson, 1978). La politesse est utilisée quotidiennement dans la langue et les stratégies associées à la politesse sont indispensables à l'atteinte des objectifs dans une interaction linguistique.

3.4.6.1 Le modèle de Brown et Levinson

Un des modèles les plus influents de la politesse a été développé par Brown et Levinson (1978, 1987). Ce modèle est basé sur la notion de « Face » proposée par Goffman (1967).

La Face est la valeur positive que l'individu attribue l'image de lui-même et qu'il essaie de maintenir dans les interactions avec les autres.

Brown et Levinson font la distinction entre la «face positive », (proposer une image de soi valorisante) qui se manifeste dans l'expression de la convivialité ou par l'approbation et la « face négative », qui implique le détachement et un besoin de liberté personnelle (défendre son territoire). Tout l'art de la bonne relation, dans le cadre d'une interaction linguistique, est dans le jonglage entre face positive et face négative. En d'autres termes, il faut respecter les attentes des interlocuteurs sans porter atteinte à leur face tout en préservant la sienne. Il faut éviter les FTA (Face Threatening Acts ou actes menaçants pour la face d'autrui ou de soi-même)

Les ordres directs, les demandes, les excuses et les insultes sont des exemples de FTA. La politesse linguistique est un des moyens d'adoucir les FTA.

3.4.6.2 La politesse et l'AL de la demande

Ainsi pour Brown et Levinson les demandes sont des FTA, et appellent donc à une action compensatrice. La requête est risquée pour la face car elle porte des enjeux élevés pour les interlocuteurs. L'émetteur sollicite une ressource du destinataire (pour des gains sociaux) et prête le flanc à un éventuel refus (les faces des deux protagonistes sont donc potentiellement menacées).

Nous avons vu dans le tableau de Trosborg (figure 3.4 plus haut) qu'il existe une large gamme de moyens linguistiques pour réaliser la requête. Toute la subtilité réside dans le choix de la stratégie appropriée (qui dépend du contexte, de la relation des locuteurs, etc.). Dans les emails professionnels, nous allons retrouver la plupart des cas de requêtes (l'ordre impératif restant cependant assez rare) afin de demander des informations, de solliciter l'accomplissement d'une tache ou de poser une question qui débutera une éventuelle résolution de problème. Nous allons donc examiner l'AL requête plus en détail.

3.5 La Requête

De nombreuses en études en pragmatique ont abordé l'acte de la requête sans doute car c'est un FTA comportant une menace élevée pour le destinataire mais également pour l'émetteur, mais aussi car c'est un acte très fréquent dans la vie courante et dans le contexte professionnel. En outre sa variabilité linguistique (directe/indirecte) et ses diverses stratégies de réalisation en font un sujet complexe. Nous allons nous concentrer sur la requête car dans notre étude sur la résolution de problème par email dans les projets, l'identification des demandes est une étape importante.

3.5.1 Requête et Demande

L'acte de la demande a été largement étudié dans le domaine de la linguistique théorique (Searle, 1969), et pragmatique interculturelle et interlangue (Blum Kulka, 1989). Un défi important dans la détection et l'identification des requêtes et demandes est qu'elles n'ont pas toujours l'air de ce qu'elles sont. De plus les emails sont connus pour contenir une grande proportion d'AL indirects (Hassell et Christensen, 1996). La difficulté augmente par la prise de compte de l'historique de la conversation et de la relation entre les locuteurs.

D'un point de vue pragmatique, une demande est un AL directif dont le but est d'obtenir de l'auditeur « d'effectuer une action dans des circonstances dans lesquelles il n'est pas évident que il / elle l'aurait effectuée dans le cours normal des événements » (Searle, 1969).

Par l'introduction d'une demande, le locuteur estime que l'auditeur est en mesure d'effectuer une action. Nous avons vu qu'il y avait deux grandes stratégies de requête en fonction du niveau de l'interprétation (de la part de l'auditeur) nécessaires pour comprendre l'énoncé comme une demande : La requête directe et la requête indirecte. La requête peut être mise en exergue, tournée soit vers l'émetteur (Puis-je faire X?) ou l'auditeur (Pouvez-vous faire X?). Une demande directe peut utiliser un impératif, un performatif, une obligation ou exprimer le vouloir ou le besoin. La forme la plus socialement brutale étant un ordre. Une demande indirecte peut utiliser des questions quant à la capacité, la volonté, etc. de l'auditeur de faire l'action ou utiliser les déclarations à propos de la volonté (le désir) de l'émetteur de voir l'auditeur faire x.

Traditionnellement il existe un débat entre les points de vue linguistiques et pragmatique sur la question et la requête. Une question est plus une demande d'information (demande d'un dire (Benveniste, 1966)) et une requête pragmatique selon Searle est plus une demande d'action (demande d'un faire).

3.5.2 Lien avec la résolution de problème dans notre approche

Dans la présente étude, nous avons limité notre recherche à l'analyse de l'acte de la requête/demande dans les séquences de résolution de problèmes. Nous avons vu que les actes illocutoires ont différentes modalités énonciatives: Il n'y a pas une relation univoque entre un acte de langage et sa réalisation linguistique. La requête indirecte en tandem avec la politesse sont souvent utilisées en contexte d'entreprise afin d'adoucir le FTA sous-jacent.

Notre approche suivant résolument une analyse de discours, nous travaillons sur des propositions qui peuvent être interprétés correctement en tenant compte à la fois du contexte discursif (qui constitue le contenu de l'e-mail) et le contexte de "énoncé" (à savoir, la situation du dialogue). Les marqueurs linguistiques seuls ne sont pas toujours suffisants pour indiquer ce qui est réellement fait dans une situation de communication médiatisée par ordinateur. Enfin, pour nous, un énoncé grammatical correspond à un seul acte de parole comme dans le tableau suivant que nous avons adapté à partir de la littérature (Trosborg,1995) (Blum Kulka, 1989) pour notre propre étude.

Type de Requête	Forme linguistique	Exemples
Requête directe	Impératif	Faites X
	Performatif	Je vous demande de faire X.
	Expression de vouloir ou de besoin.	J'ai besoin/ Je veux que vous fassiez X
	Obligation	Vous devez faire X
Requête indirecte	Questionnement sur la : Capacité de l'interlocuteur	Pouvez-vous faire X? Pourriez-vous faire X?

	Volonté de l'interlocuteur	Voudriez-vous faire X?

Figure 3.5 : Grille d'analyse requête

3.6 Analyse du discours et email

Depuis les premiers travaux sur la CMDA, les linguistes et les philosophes se sont questionnés sur la pertinence et les différentes manières d'étendre la théorie des actes de langage au domaine de la conversation et du discours.

Un des effets pragmatiques majeurs causé par le courrier électronique est celui des dialogues par email (en outre si les personnes sont simultanément devant leur client email, cela peut dériver vers une forme de messagerie instantanée quasi synchrone avec abandon des formules de politesse, (et des signatures, salutations), et ajout d'émoticônes, écriture dans les corps des messages précédents.

Vanderveken (2001) présente une version claire de cette proposition lorsqu'il écrit : « les locuteurs exercent leurs actes illocutoires dans les conversations entières où ils sont le plus souvent dans l'interaction verbale avec d'autres intervenants qui leur ont répondu et effectuent à leur tour leurs propres actes de langage avec la même intention collective de poursuivre avec succès un certain type de discours. (..). Il se compose, en général, des séquences ordonnées de déclarations faites par plusieurs orateurs qui ont tendance, par leurs interactions verbales pour atteindre des objectifs discursifs communs (comme discuter d'une question, décider ensemble comment réagir face à une certaine situation, faire une négociation, se consulter ou plus simplement pour échanger des salutations parler de soi) (..) Pour plus de commodité terminologique, je vais appeler ces séquences ordonnées des actes de discours des conversations. " . Ce sont précisément ces actes là en lien avec la résolution de problème qui nous préoccupent dans cette étude.

3.6.1 Relation/Rôle et lien avec l'organisation en milieu professionnel

Nous avons vu plus haut qu'il est impossible de ne pas communiquer, un silence est une réponse. Dans un cadre d'entreprise, ne pas répondre à une demande par email de

son supérieur peut être vu comme un désaccord ou même une insubordination. De plus la réponse (ou son absence) déterminera tout le cours ultérieur de la conversation.

On retrouve dans le courrier électronique l'équivalent des tours de parole de l'analyse conversationnelle (Kerbrat-Orecchioni, 2008). Les mots écrits et la façon dont ils sont enchainés forment la conversation et portent deux aspects : le contenu et la relation.

Il faut aller plus loin que le modèle émetteur/récepteur et le simple codage/décodage de l'information. Ce qui importe est non seulement ce qui est dit, mais la façon de le dire et le contexte dans lequel il est dit. La manière de percevoir un message dépend éminemment de la relation entre les participants. Cela peut modifier le contenu effectif d'un message, par exemple dans un projet quelqu'un qui agira de manière autoritaire selon qu'il en a la légitimité (hiérarchique conventionnelle) ou pas ne se verra pas offrir le même type de réponse ou d'aide. Le fil entier d'une conversation (par email également) peut se décider selon le ton du message initial et la relation existante entre les utilisateurs. De même la séquence d'échange entre les participants d'un projet définira (ou redéfinira) peu à peu leur relation.

Nous avons vu que la particularité de la plupart de la CMO (si l'on exclut la vidéo conférence) et du courrier électronique est qu'elle véhicule mal le non verbal (expression faciale, geste, mouvements, etc..). Cela peut avoir un impact important sur la relation (et son évolution) entre les participants d'une conversation, les mots seuls expriment mal le ressenti et peuvent donner lieu à de nombreux malentendus organisationnels. C'est la différence qu'expose Watzlawick (1972) entre la communication digitale (contenu) et analogique (relation). Cette distinction est notable, elle explique en particulier pourquoi des personnes avec des points de vue opposés ne parviennent pas à une véritable compréhension partagée, même quand ils comprennent le contenu de la position des autres. Cela peut se résoudre en objectivant les questions, les arguments et les idées grâce aux outils de cartographie de dialogue.

Pour aller plus loin cela pose la question d'une réalité objective d'une situation de communication et de la multitude des points de vue. Fort heureusement dans notre cadre de la connaissance en mémoire de projet, le contexte et les intervenants sont essentiels. La connaissance existe en contexte (en contexte au sens large

organisationnel, technique et humain) aussi les points de vus différents de la réalité sont à prendre en compte.

3.6.2 Rôle/Subordination/ Domination

Selon Watzlawick (1972) tous les échanges communicationnels sont symétriques ou complémentaires, en fonction de la relation entre ceux qui sont impliqués, à savoir si elle est basée sur l'égalité des participants ou sur les différences entre eux. Suivant cette approche, les interactions symétriques sont basées sur la minimisation des différences entre les intervenants et relations complémentaires sur la maximisation des différences. Aucune n'étant préférable à l'autre en soi, c'est simplement un mode d'interaction.

C'est cependant primordial dans notre travail, car si un manager considère ses employés comme des pairs, il ne formulera pas ses requêtes de la même manière que si l'équipe prend une attitude de soumission (échange complémentaire). Par extension, il sera essentiel dans notre étude de correctement qualifier les relations de subordination/soumission officielles (issues de la hiérarchie, ou d'un contrat fournisseurs/donneur d'ordre) car les requêtes et demandes prendront des formes différentes. Pour éclaircir ce point, si un manager écrit à un employé pour signaler une erreur, il faudra y voir une requête indirecte de demande de corrections (même si dans la forme linguistique cela n'est pas évident).

3.7 Conclusion

Nous allons essayer d'appliquer des techniques d'analyses pragmatiques et conversationnelles pour en extraire une structuration des messages professionnels (emails dans le cadre d'un métier). Cette structuration va nous aider à déterminer **quel rôle joue le mail dans la traçabilité et la mémoire de Projet**. La problématique consiste donc, à l'aune de la Mémoire de Projet, à l'identification, l'analyse des parties récurrentes des messages, leur typage en utilisant la théorie des actes de langages, et enfin à leur organisation afin de définir des accès aux connaissances et pratiques évoquées dans ces messages.

La linguistique s'est intéressée à la communication médiatisée sous divers aspects des messages (Forum, BBS, Listes de diffusion, Blog, Chat, etc..) principalement en utilisant des méthodes d'analyse du discours. La plupart du temps les codages et les analyses sont faits manuellement.

Certains travaux ont eu lieu sur l'email (par exemple Baron(1998), Herring (2002,2004)) mais davantage centrés autour de ses caractéristiques linguistiques « pures » (différences avec l'écrit et le parlé, style syntaxique) par exemple il existe beaucoup d'études sur les traits linguistiques de l'email (modaux, champs lexicaux, abréviation, densité lexicale, etc..), mais peu sur les aspects sémantiques, pragmatiques et relationnels. En effet dans une communication médiatisée, non seulement les techniques statistiques de détection des occurrences de termes ne peuvent pas être effectuées (de par la brièveté des messages), mais il existe une dépendance sémantique très forte entre les différents messages déjà échangés. Enfin la plupart des corpus emails utilisés dans les articles sont très spécifiques (de faible taille et généralement issus du milieu académique et/ou étudiant).

La pragmatique de la communication et la théorie des actes de langages de Searle a été relativement peu utilisée dans l'analyse des emails. L'analyse conversationnelle a souvent lieu sur des discours, des livres, des dialogues (comptes rendu de réunion).

La question se pose de savoir dans quelle mesure les actes illocutoires (intentions du locuteur) passent-ils convenablement via le canal de l'email. De même la séparation spatiale et temporelle peut créer des distorsions par rapport à une conversation en face à face. La co-construction du discours qui s'effectue dans ce cas-là est-elle vraiment similaire, et surtout quel sera le moyen technique le plus adapté pour en retirer des informations utiles pour l'ingénierie des connaissances, et en particulier la Mémoire de Projet.

Nous avons vu dans le chapitre 1 que pour une entreprise, la Mémoire de Projet est une trace du capital Connaissance acquis pendant la réalisation d'un projet. Les décisions, les participants, les taches, les actions, les stratégies, la logique de conception : tout cela fait partie de la Mémoire de Projet (MP) et puise sa source dans des documents, des discussions, des prototypes. L'email est rarement utilisé en tant que tel du point de vue de l'ingénierie des connaissances (souvent pour les pièces

jointes et pour un aspect social, de coordination et de planification), pourtant il recèle des informations sur la relation entre les participants, l'organisation émergente qui se manifeste pendant le projet (dans l'action et dans un certain contexte) , les démarches de résolutions collectives de problème et il est légitime de vouloir trouver une méthode pour exploiter ce potentiel et le stocker.

Notre approche est basée sur l'idée que le contenu d'un courrier électronique dans un fil (thread) peut se résumer en un certain nombre d'actes de langage prédéfinis. Une des difficultés est de trouver la grille de lecture adaptée, c'est-à-dire le type d'informations utiles pour la mémoire de projet présentes dans l'email (par exemple les taches, les contraintes, les suggestions, etc..) et ensuite de trouver une codification en actes de langage qui pourrait nous amener à détecter ces éléments.

Nous allons nous focaliser plus finement sur la résolution de problèmes et la logique de conception et donc notre grille d'analyse a été conçue à cet effet. Notre analyse des actes de discours sur la résolution de problème va tenter de localiser les demandes (requête) et les possibles solutions (réponses). En premier lieu il s'agira des actes de discours qui aident à localiser une demande dans un message. Ensuite, nous étudierons l'organisation des messages apparentés dans le fil afin d'identifier la solution proposée (si elle existe) à la demande. On s'intéressera aussi à l'ensemble des indices verbaux ou non contenus dans la situation. Donc, nous combinerons analyse pragmatique et l'analyse textuelle du contenu et nous relierons ces analyses au contexte du projet (les compétences et rôle des expéditeurs et des récepteurs, les phases du projet, les livrables, etc.) afin de garder une trace de l'intention et de donner un sens aux interactions entre les acteurs.

Il est également envisageable d'essayer d'identifier/typer des mouvements plus larges (sur un fil de mail) ou même sur plusieurs fils de mail si l'on veut détecter par exemple l'apparition de conflits ou la prise de décision.

Mais nous allons pour cela avoir besoin d'étudier les traitements du langage naturel et particulièrement ceux gravitant autour de l'email, des actes de langage. Ce sera l'objet du chapitre suivant.

Chapitre 4. EMAIL ET TRAITEMENTS DU LANGAGE NATUREL

4.1 Présentation

Dans le chapitre précédent, nous avons abordé la communication médiatisée, et plus particulièrement ce qui est au centre de notre étude : l'Email. L'objet de notre étude est de déterminer dans quelle mesure il est possible d'extraire des traces de connaissances pertinentes pour la mémoire de Projet depuis un corpus d'email professionnel. Il s'agit donc de traiter un volume important de messages, donc de données textuelles. Pour des recherches algorithmiques et méthodologiques, il est possible de travailler manuellement sur un échantillon de données, mais il est bien entendu que notre démarche repose sur des outils informatiques qui sont les seuls en mesure de traiter le volume de donnée réel de manière efficace. En effet, un corpus représente des milliers de messages, donc des dizaines de milliers de phrases à examiner. Nous allons donc nous intéresser dans ce chapitre aux approches en termes de modèles et d'algorithme qui ont trait au langage naturel, plus particulièrement écrit et bien entendu aux recherches autour des messages électroniques.

Dans ce chapitre, nous présentons dans un premier temps, les principaux éléments à analyser dans un texte, à savoir mot, phrase et document. Nous décrivons certaines

approches permettant les analyser. Dans un deuxième temps, nous présentons les spécificités de ces analyses sur les e-mails. Enfin, nous montrons les principaux concepts que nous avons retenus dans notre approche.

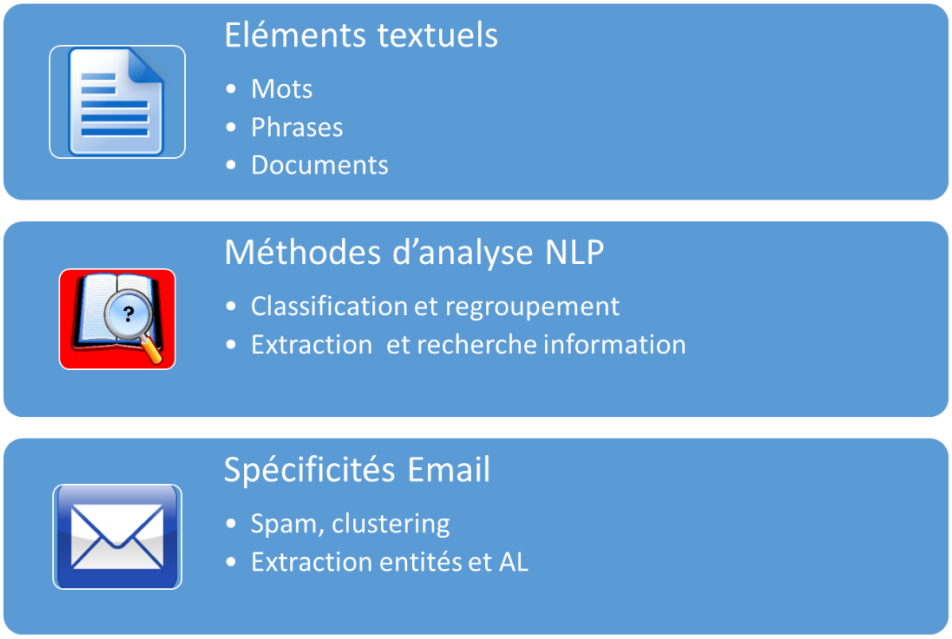


Figure 4.1 : Plan synthétique du chapitre

4.2 Le Traitement du langage naturel

Le traitement du langage naturel (Natural Language Processing ou **NLP**) (Collobert, 2011) est un domaine de recherche ayant une histoire assez longue et qui se trouve à la frontière de la linguistique computationnelle, de l'intelligence artificielle, et du Textmining. Les systèmes d'information actuels permettent aux entreprises de traiter de grand ensemble de données (et même de très grands en centaine de To avec l'avènement du Big Data), du moment qu'elles sont plus ou moins structurées et peuvent être analysés avec des bases de donnée (y compris NoSQL[7]).

Cependant ce n'est pas le cas des données textuelles qui sont par essence non structurées et très difficiles à traiter. A la différence du Datamining ou les informations que l'on cherche à découvrir sont implicites, ou cachées, en Textmining les

[7] NoSQL : « Non SQL ou Not Only SQL » base de donnée non relationnelles distribuées pouvant accueillir de très grands volumes de données par exemple Hbase (Google), Cassandra(Facebook), CouchDB, etc..

informations sont claires et explicites dans le texte, seulement elles ne peuvent pas être assimilées directement par une machine.

Le NLP (Jurafsky, 2000) vise à convertir le langage humain vers une représentation formelle qui facilite sa manipulation informatique. Le rêve ultime serait bien sûr de pouvoir donner une œuvre de Victor Hugo à un logiciel est d'en tirer une structure informatique qui décrive complétement et sans ambigüité le sens du texte et la pensée de l'auteur. La compréhension complète (sémantique, pragmatique) du langage naturel est encore un objectif assez lointain cependant diverses techniques ont émergé de ces recherches. Les scientifiques ont travaillé sur des sous problèmes plus abordables, en extrayant des représentations plus simples qui décrivent des aspects restreints de l'information textuelle.

Les applications actuelles vont de la traduction automatique (Google Translation par exemple), les résumés automatiques, la classification ou la recherche de texte et les interfaces homme-machine (Collobert, 2008).

4.3 Analyse de texte par ordinateur

Pourquoi est-il si difficile d'analyser du texte ? C'est que le langage humain est peu structuré, abstrait, parfois vague, ambigu (Smith 2011). Les mêmes relations et les concepts sont présentés de multiples façons. La compréhension du sens suppose l'utilisation de connaissances du domaine, implicites ou du contexte (y compris social). Le raisonnement peut être nécessaire et surtout d'un point de vue informatique les données peuvent être décrites avec un très grand nombre de caractéristiques. Il est donc important de définir des structures permettant de représenter des éléments du texte.

4.4 Représentation des textes

Le langage humain peut s'appréhender à différentes granularités. Le mode de représentation dépendra de la tâche à effectuer. Dans notre étude, nous n'aborderons pas les langages à base d'idéogrammes comme le chinois mandarin ou le japonais par exemple.

On peut représenter formellement à un bas niveau le langage comme des chaines de caractères assemblées (Zaragoza, 2013) :

- Alphabet $\Sigma=\{ a,b,c\}$
- Chaine de caractère s = aabbabcaab
- Ensemble de toutes les chaines possibles: $\Sigma^* = \{a,b,c,aa,ab,ac,aaa,\ldots\}$
- Langage (Formel): $L \subseteq \Sigma^*$

Un langage formel étant basé sur une grammaire avec des règles de production (Chomsky, 1972). Dans le langage naturel nos mots font office de caractères et nos phrases sont constituées de chaines de « mots ». Le mot est l'unité fondamentale du texte, et c'est à partir des mots que se construit le sens. Une phrase est l'unité d'action du texte, elle-même assemblée en paragraphe, chapitre, et enfin document.

Le langage naturel est multi niveau et son étude recouvre plusieurs domaines. Au niveau des mots, il y a la morphologie linguiste qui étudie leur forme ou leur flexion, (et aussi la phonologie), et la sémantique lexicale qui étudie leur sens. Au niveau de la combinaison des mots en phrase (via une structure grammaticale), il y a la syntaxe. On retrouve la sémantique au niveau des phrases, et enfin à un niveau plus vaste du document, des paragraphes et coréférences, il y a la pragmatique et l'analyse du discours.

Ces différentes disciplines ont leur outils et modèles, et à notre échelle, on utilisera les grands modules suivants :

- Lexique
- Syntaxe/Grammaire
- Sémantique
- Pragmatique

Cette division reste volontairement simple et provient des outils informatiques traditionnellement employés en NLP (Nadkarni, 2011). En effet lorsque l'on analyse du contenu textuel par ordinateur, il est d'usage d'utiliser un lexer (analyseur lexical pour découper le contenu en token), puis un parser (pour tenter de retrouver une structure grammaticale) et enfin des traitements de plus haut niveau pour la

sémantique ou la pragmatique. Nous allons revoir en détail les traitements possibles selon le niveau étudié.

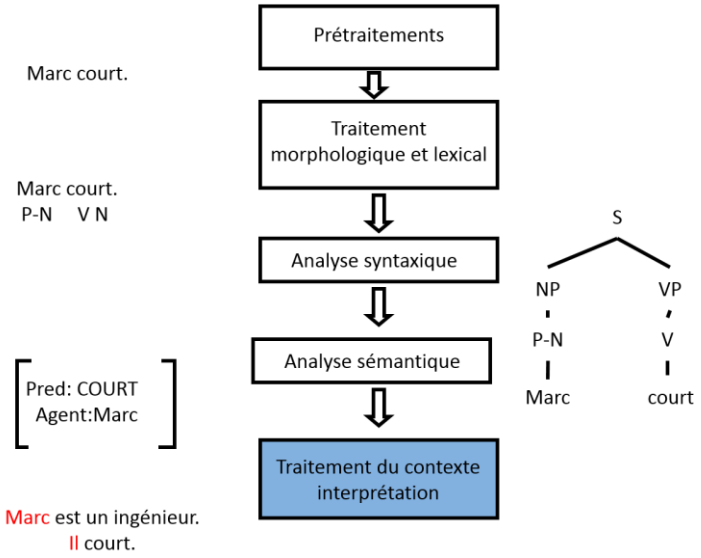


Figure 4.2 : Schéma général des traitements en NLP

Nous présentons par la suite, les différentes techniques qui ont été définies pour traiter les différentes parties d'un texte : mots, phrases et documents.

4.5 Les Mots :

4.5.1 Découpage

Evoquons brièvement ici le niveau caractère, c'est-à-dire l'analyse d'un texte en étudiant les chaines de 1, 2,3,.. caractères qui peut être utile dans certains traitements très spécifiques (détection de spam (Cormack 2007) ou de plagiat (Stamatos 2009) par exemple) mais qui se révèle trop simple et auquel on préfère généralement le traitement au niveau mot.

Afin d'analyser un texte, on le divise tout d'abord en mots (ou token). Cette phase emploie des caractères courants (espace, signe ponctuation, paragraphe) pour séparer les mots et ne pose généralement pas de problème dans les langues qui nous concernent ici.

Les mots résultants soulèvent déjà des problématiques connues dans leur analyse :

- La polysémie (un mot avec des sens différents (par exemple avocat) pour lequel des techniques existent pour lever les ambiguïtés (Ide, 1998)
- L'homonymie/homographie (par exemple « les poules du couvent couvent »)
- La synonymie (mot différent avec des sens identiques) et l'hyponymie/hyperonymie (« néon est une lampe », « fraise est un fruit »)

On peut également citer Wordnet[8] un thésaurus anglo saxon (il existe une version française en cours de développement) très important dans la communauté NLP. Il recense les mots en les regroupant en synsets (groupe de mot interchangeables véhiculant le même sens) et les relations d'hyperonymie.

4.5.2 Fréquence et répartition

Il est connu de longue date que les fréquences de mots suivent l'inverse d'une loi de puissance (Zipf, 1932). La fréquence d'occurrence $f(n)$ d'un mot est liée à son rang n dans l'ordre des fréquences par une loi de la forme $f(n) = \frac{K}{n}$ où K est une constante. Cela a d'abord était établi sur l'anglais il a été montré plus tard (Mandelbrot, 1982) (Li, 1992) que l'on pouvait trouver des lois similaires sur des textes aléatoires. La distribution des fréquences en elle-même nous amène peu d'information et semble même être une caractérisation pauvre pour distinguer le langage naturel. Les mots les plus fréquents amènent peu de sémantique. Nous reviendrons sur cet aspect lors de la description du Vector Space Model.

4.5.3 Prétraitements

Deux autres aspects sont essentiels dans le prétraitement des textes, la troncature aux racines (ou racinisation, en anglais stemming) et la suppression des mots vides (stop words removal).

Le stemming (Porter, 2001) consiste à supprimer les suffixe et préfixes des lemmes pour ne conserver que la racine. Cela améliore les résultats en recherche d'information

[8] http://wordnet.princeton.edu/

surtout sur les requêtes courtes (Kantrowitz, 2000). Le stemmer le plus utilisé est celui de Porter.

La suppression des mots vides consiste à ne pas indexer et à enlever des traitements certains mots très communs car ils n'apportent rien. Par exemple « le », « la », « de », « du », « ce »… Là aussi en recherche et classification de documents cela peut améliorer les résultats (et en tous les cas alléger l'espace de recherche) (Silva, 2003 ; Wilbur, 1992)

Nous verrons également que le prétraitement d'un courrier électronique a ses spécificités (partie répétées, réponses incluses, signature, etc.)

4.6 Les Phrases

Il existe différents traitements au niveau de la phrase en fonction des tâches à accomplir. Les plus fréquents sont les n-grammes et l'étiquetage morpho-syntaxique (Part Of Speech tagging ou POS).

Les n-grammes sont une séquence contiguë de n-mot issus d'un même texte. On peut voir à l'œuvre une analyse basée sur les 1,2..,5 grammes dans le projet Google Book[9]. Les modèles n-grammes sont employés dans les modèles de langages probabilistes (c'est-à-dire permettant de prédire la probabilité d'une séquence à partir des précédentes) souvent basés sur les modèles de Markov (Martin, 2000). Ces modèles étaient utilisés entre autre pour la traduction automatique (Brown 1990), l'assistance à la génération de réponse emails automatisées (Bickel, 2005) et la reconnaissance de la parole.

Un POS est un logiciel qui assigne à chaque partie du discours une étiquette grammaticale (un tag) comme nom, verbe, adjectif, etc.

Les POS ont été implémentés avec de nombreux modèles (Modèles de Markov Cachés, Perceptron, Support Vector machine, Maximum entropy, etc., un état de l'art complet est disponible sur ACLweb[10]).

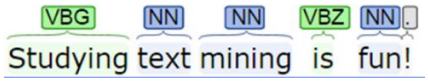
[9] https://books.google.com/ngrams

[10] http://aclweb.org/aclwiki/index.php?title=POS_Tagging_%28State_of_the_art%29

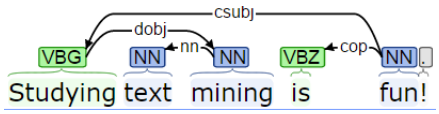
Actuellement l'un des plus efficace et plus employé est celui de l'université de Stanford[11]. Les POS sont très utiles car ils servent d'entrée (pour extraire des paramètres) à des systèmes de plus haut niveau (pour extraire des relations sémantiques (Kambhatla, 2004), pour analyser les sentiments, créer des arbres de dépendances, etc. (Jurafsky, 2000)). Nous les retrouverons fréquemment dans les traitements NLP sur l'email et/ou la pragmatique.

Voici un exemple tiré de H. Wang[12] sur la phrase « Studying Text mining is fun ! »

Le Part Of Speech tagging donnera :



Et l'arbre de dépendance sera :



A noter que dans notre étude, la phrase revêt un caractère important car elle est l'unité qui portera les actes de langage.

4.7 Le Document

Nous avons vu les traitements et modélisations possibles au niveau mot et phrase. Il faut aussi considérer un ensemble de documents. Dans notre étude, cela sera un ensemble de messages électroniques.

Le Vector Space Model (VSM) a été développé par G. Salton et son équipe (Salton, 1975). Il est à la base des concepts mis en œuvre dans la plupart de moteurs de recherche (Mac Candless, 2010).

[11] http://nlp.stanford.edu/software/tagger.shtml

[12] https://www.cs.virginia.edu/people/faculty/hwang.html

Le principe du VSM est de représenter chaque document comme un vecteur dans un espace vectoriel dont la dimension est la taille du vocabulaire. Les vecteurs proches de cet espace sont sémantiquement similaires. Dans la cadre de la recherche, la requête utilisateur est assimilée à un pseudo document et donc à un vecteur de cet espace et les documents correspondants sont présentés par ordre de similarité décroissante avec cette requête.

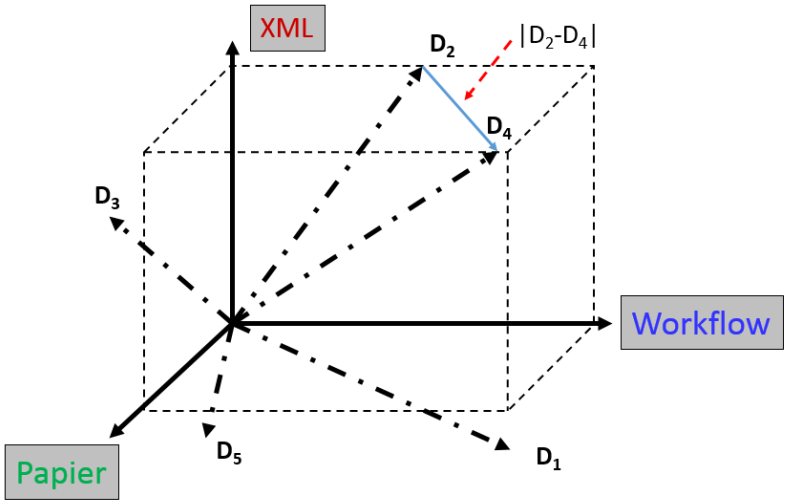


Figure 4.3 : Projection de documents dans le VSM avec des exemples de termes sur les axes

4.8 VSM, TFD-IDF et similarité

La recherche de document avec le VSM utilise l'hypothèse du sac de mot (BOW Bag of Words) (Salton, 1975). Celle-ci soutient qu'un vecteur colonne dans la matrice terme-document (Baeza-Yates, 1999) capture (d'une certaine manière) un aspect sémantique du document : de quoi parle le document. Cet aspect sera primordial dans notre analyse des messages emails. L'approche BOW ne conserve pas l'ordre ni la grammaire.

Plus formellement la matrice terme-document simple associe à chaque ligne terme sa fréquence d'apparition dans la colonne document.

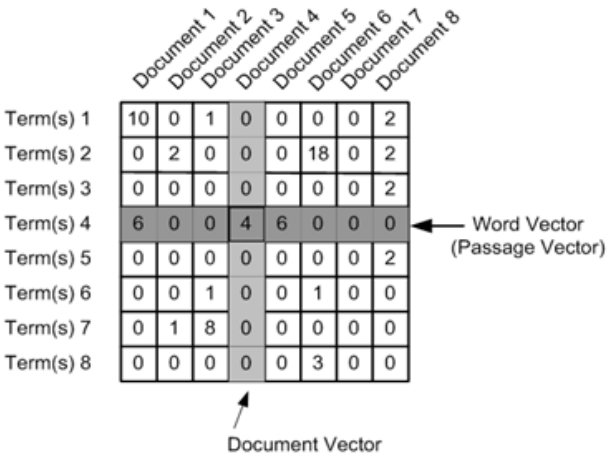


Figure 4.4 : Matrice terme-document (d'après B. Rose[13])

Trois facteurs sont à mettre en lumière dans cette première approche :

- La fréquence des termes (term frequency ou tf) dans un document marque leur importance (avec les réserves sur le stop word removal)
- Un terme trop fréquent sur un grand nombre de documents se révèle au final peu indicatif, d'où la mise en place d'une méthode de pondération basée l'inverse de la fréquence par document (inverse document frequency ou idf) ((Sparck Jones, 1972).
- Les documents longs ont tendance à avoir des scores plus important car contenant plus de mots et de répétitions. Cela est corrigé en normalisant la pondération par rapport à la taille des documents.

Au final l'approche TF-IDF se présente ainsi :

Si on note $A = (a_{ij})$ la matrice terme-document avec la pondération TF-IDF avec m termes et n documents

$$a_{ij} = tf_{i,j} \cdot idf_i$$

Avec

[13] http://brandonrose.org/clustering

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

où $n_{i,j}$ est le nombre d'occurrence du terme i dans le document j.

et

$$idf_i = 1 + \log(\frac{n}{1 + doc_i})$$

où doc_i est le nombre de document contenant le terme i.

Des variantes existent pour le lissage des poids (par exemple en utilisant l'entropie des termes et en prenant le logarithme des tf (Dumais,1991)) mais nous resterons dans l'approche ci-dessus similaire à l'implémentation dans le logiciel Lucene.

Une fois cette matrice terme document établie dans le VSM, il est alors possible de mettre en œuvre des mesures de similarité entre documents (ici d_i et d_j). Une distance euclidienne est possible du type :

$$dist(d_i, d_j) = \sqrt{\sum_t [tf_{t,i} idf_t - tf_{t,j} idf_t]^2}$$

Mais elle pénalise les documents longs et montre mal comment les vecteurs se recouvrent. Aussi a été mise en place la Cosine Similarity qui prend en compte l'angle dans le VSM.

$$sim(d_i, d_j) = \frac{\overrightarrow{d_i}.\overrightarrow{d_j}}{\|\overrightarrow{d_i}\|.\|\overrightarrow{d_j}\|}$$

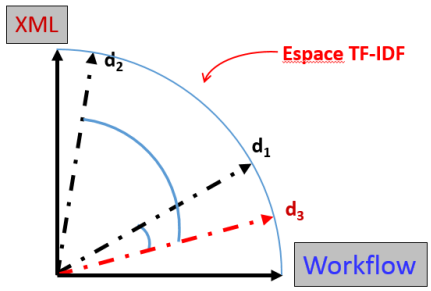


Figure 4.5 : Cosine Similarity entre vecteurs dans l'espace VSM/TFIDF

L'approche VSM/TFIDF est assez robuste et largement usitée, cependant elle ne couvre pas les problèmes d'homonymie et polysémie. Cela a donné lieu à des évolutions (décomposition en matrices singulières de la matrice TFIDF par Latent Semantic Analysis (LSA) (Deerwester, 1990)) ou encore Latent Dirichlet Allocation (LDA) (Blei, 2003) un modèle génératif qui prend mieux en compte la polysémie. Ces modèles plus lourds à mettre en œuvre n'entreront pas dans la cadre de cette étude.

Cependant, la recherche d'Information peut fournir des techniques de similarité intéressantes pour l'analyse des messages qui ne sont pas aussi complets qu'un document.

4.9 Recherche d'Information et mesures

Nous avons vu dans le chapitre 2 que la traçabilité utilise l'IR (Information Retrieval) pour retrouver des liens entre données textuelles et artefacts de produits (par exemple codes informatiques (Baeza-Yates 1999)).

L'IR permet de rapprocher des documents par leur similarité, ou par rapport à une requête et d'en extraire des informations. Il faut ensuite mesurer la validité des résultats et la performance de la solution. Pour ce faire on utilise trois indicateurs très simples : la précision (precision), le rappel (recall) et la F-mesure (moyenne harmonique des deux) (Manning, 2008). Ils sont définis à partir des résultats. En IR, on doit retrouver des informations et vérifier qu'elles sont correctement classées (ou étiquetées). En supposant une classification binaire (positif/négatif), les « true

positive » (tp) sont les documents bien classés en positif, les « false positive » (fp) sont les document classés en positifs mais en fait négatifs, et symétriquement pour les « true negative »(tn) et « false negative » (fn). On définit alors :

$$precision = \frac{tp}{tp + fp} \quad recall = \frac{tp}{tp + fn}$$

Et enfin

$$F - Measure = 2 \times \frac{recall \times precision}{(precision + recall)}$$

On voit qu'il est aisé d'atteindre 1 en recall (100%) en retournant tous les documents en réponse à une requête. Par conséquent, le recall seul est insuffisant et il faut mesurer le nombre de résultat non pertinents via la précision. Un recall de 1 signifie que tous les résultats pertinents ont été récupérés, mais il pourrait aussi y avoir des résultats non pertinents. Si la précision est de 1, cela implique que tous les résultats récupérés sont pertinents, mais il y pourrait y avoir des résultats pertinents non récupérés. La F-measure combine les deux indicateurs en une moyenne.

4.10 Traitements autour de l'email

4.10.1 Principales application en NLP

Les traitements textuels issus du Textmining et du NLP peuvent être regroupé en grande familles :

- Extraction d'information pour les utilisateurs : Recherche de document, recherche d'information, résumé de texte, etc.)
- Vérification de similarité de document (catégorisation de texte, regroupement (clustering) de documents (avec les K means par exemple (Rocchio, 1966)), attribution d'auteur, etc.)

- Extraction d'informations structurées (Extraction d'entités nommées (nom de personnes, d'organisation, d'email, d'adresse, etc..), de relations entre entités, apprentissage de règles, etc.)

Les applications pratiques sont nombreuses :

La catégorisation de texte se retrouve dans les filtres d'emails ou la labellisation automatique dans les bibliothèques professionnelles (Miller 2005), (Sebastiani, 2002), le clustering de documents a été appliqué au regroupement d'articles de presse (Bouras, 2010). La recherche d'information est devenue omniprésente avec les moteurs de recherche sur internet (Brin, 1998). Egalement dans l'identification du cœur des disciplines dans le domaine des systèmes d'information (Larsen 2008), la gestion de la relation client (Cheung 2003), et la découverte de topics (Pons-Porrata 2007). Dans des objectifs différents, l'identification et la classification des sentiments ou opinions à partir de media sociaux (Go, 2009) ou journaux professionnels pour investisseurs (Tetlock, 2008).

L'email étant au cœur de notre démarche, nous allons passer plus spécifiquement en revue les traitements textuels autour de celui-ci.

4.11 Email

Nous avons vu dans le chapitre précédent que le courrier électronique était un message textuel avec :

- Un entête (ou méta données comprenant l'adresse de l'émetteur, celles des destinataires, la date, et des données informatique sur les serveurs emails et le format).
- Un sujet.
- Un corps (des mots ou phrases avec/sans salutation et signature).
- Des éventuelles pièces jointes (des fichiers joints comme des images ou des documents).

En outre, il s'insère dans un fil de conversation (thread) regroupant plusieurs messages liés par des questions/réponse/transferts assimilable à une forme de discours.

Cependant l'email présente des caractéristiques spécifiques du point de vue du Textmining (Dalli, 2004).

- Des messages courts (habituellement 2 à 500 mots (après suppression des parties répétées))
- Il n'obéit souvent pas aux formes stylistiques et grammaticales usuelles (abréviation, ponctuation, erreur de frappe)
- Il est un mélange de style formel et informel (proche du SMS) pouvant évoluer au cours du temps et de la relation des usagers. (le style informel augmente la proportion de mots déictiques et donc la nécessité de la prise en compte du contexte)
- Les mails font souvent partie d'un fil (30% en moyenne dans l'étude de Fisher) et contiennent des parties répétées (les mails précédents). 87% des messages reprennent au plus 3 des messages précédents (Fisher and Moody, 2001)

La taille courte longueur des messages emails entraine peu de cooccurrence pour les mesures de similarité, aussi les méthodes standards de Machine Learning (Phan, 2008), statistiques ou probabilistes ont généralement du mal à atteindre de bonnes performances en raison de la dispersion des données (Aggarwal, 2012). Afin de pallier à ce problème d'autres approches sont nécessaires, on peut en distinguer deux principales :

- D'abord étendre la représentation des textes, c'est-à-dire exploiter différents aspects des phrases pour préserver l'information contextuelle. Cela peut se faire par combinaison du BOW avec par exemple des synonymes, hyperonymes, hyponymes, etc.. de Wordnet (les chaines lexicales des travaux de Stokes (Stokes 2001)). Pour illustrer ceci : avec un document concernant les voitures, la chaine lexicale typique consisterait en {véhicule, moteur, roue,

voiture, automobile, volant}. De manière similaire, dans le modèle de Yang (Yang, 2002), appliqué à la détection de nouveauté dans des documents courts (actualités), on prépare des sujets (topics) pour classer les documents, ces topics sont définis manuellement par des mots clefs, puis dans une seconde phrase, on pratique la détection de nouveauté avec des mesures de similarité à l'intérieur de chaque groupe de topics.

- L'utilisation de connaissances extérieures pour franchir le fossé sémantique (« semantic gap ») dans la représentation des textes (Gabrilovich, 2005). L'exemple de l'ouvrage d'Aggarwal est parlant : les mots « Tremblement de terre au Japon » ne contiennent aucun mot ou phrase ayant trait à la « Crise Nucléaire » mais on peut facilement s'apercevoir que ces deux évènements sont liés en parcourant des sites d'actualités. L'idée intuitive sous-jacente est qu'il faut enrichir les textes courts en exploitant des ressources extérieures pour réduire le « semantic gap ».

Nous verrons à la fin de chapitre que ces deux thèmes sont essentiels pour nos recherches.

4.12 Approches de traitements des e-mails

Les travaux en traitement de l'email sont très nombreux même s'ils sont en diminution depuis quelques années au profit des études autour des messageries instantanées ou des réseaux sociaux (Twitter, Facebook). Il y a différentes façons d'aborder la littérature sur le sujet tant elle est foisonnante, que soit au niveau des objectifs ou des méthodes. Dans le cadre de cette étude, nous nous orienterons plutôt sur un aspect historique et « niveau d'abstraction ». En effet les problématiques autour de l'email ont plus ou moins été les même depuis ces débuts, et les méthodes se retrouvent avec des ajouts dus aux évolutions technologiques (et à l'avancée de la recherche). Ce qui a changé c'est la façon d'aborder les problématiques textuelles. On distinguera les

axes suivant donc (suivant le découpage évoqué plus haut dans ce chapitre) et qui rejoint peu ou prou la chronologie des travaux :

- Syntaxe/ Lexique/Grammaire
- Sémantique
- Pragmatique

A noter que ces modalités d'approche s'emboitent souvent les unes dans les autres et que nous distinguerons la principale. Leurs objectifs allant de pair avec ces niveaux d'analyses que l'on retrouve au fils du temps :

- Recherche d'information
- Classification
- Détection d'information
- Extraction d'information
- Clustering (regroupement)

Et les méthodologies plus ou moins sophistiquées et empruntées des disciplines connexes comme NLP, Machine Learning, Datamining, Analyse probabiliste et statistique, Linguistique computationnelle, Analyse des réseaux sociaux, etc. En dernier lieu, nous nous attarderons plus longuement sur les aspects pragmatiques liés à notre étude.

4.13 Classification et détection

Il existe un nombre important de travaux sur le filtrage des emails dû à l'apparition rapide du spam (courrier non sollicité). Sur la détection du spam avec des filtres bayésiens (Androutsopoulos 2000) (Robinson 2003) ou la détection du phishing (hameçonnage pour attirer l'usager vers des sites dangereux) avec des vecteurs

caractéristiques (Fette, 2007). Un panorama complet a été fait dans Guzella (2009) et plus récemment dans Gomez (2012).

La tâche de filtrage est difficile car les techniques des émetteurs de mails douteux varient dans le temps et s'adaptent aux contre mesures, mais à l'heure actuelle ces méthodes sont souvent présentes et fonctionnent assez bien dans les logiciels grands publics (Outlook, Thunderbird, Spam assassin, Gmail, etc..). Ce sont cependant des tâches éloignées de notre problématique car elles visent à séparer les emails en quelques classes le plus souvent par des méthodes statistiques, alors que nous visons au contraire analyser un ensemble des e-mails liées à leur contexte.

4.14 Détection d'information

La détection de topics est apparue plus tard et la plupart des approches repose sur des méthodes issues du Machine Learning ou du Datamining. Li (2006) a combiné l'analyse sémantique et la détection de l'entité nommées (NER) pour la détection de topics et Surendran (2005) a étudié la gestion des topics personnels pour le courrier à l'aide de mots clefs. Certaines études se sont attachées à retrouver les émetteurs en utilisant le nombre de mots, de lignes et des mots clefs (De Vel, 2000)

On peut également détecter les comportements anormaux d'un émetteur (Stolfo, 2006) en analysant ces patterns d'envoi de mail et en détectant des variations significatives (via des réseaux neuronaux ou des SVM). Nous exploitons ce type d'approches afin d'isoler les messages relevant des objectifs visés par l'analyse ce ceux qui paraissent très généraux.

4.15 Extraction/Recherche d'information

Cependant il est possible d'utiliser les techniques de Textiming décrites plus haut pour analyser plus finement les emails (Rennie, 2000), et essayer d'en extraire des informations. Une grande partie des recherches sur l'email est consacrée à des tâches

concernant les personnes comme la reconnaissance des noms et des références (Diehl, 2006) (Minkov, 2005, 2006), l'extraction d'information de contact (Culotta, 2004), et la recherche d'expert (Balog, 2006), ou encore prédire les actions à effectuer en fonction d'un email (Dabbish, 2005).

Laclavick (2012) reprend le concept de « gazetteers » (introduit dans GATE (Cunningham, 1995)), les gazetteers étant des "dictionnaires" de mots clef (semblable aux nomenclature pour les lieux ou les organisations), la définition exacte de l'article étant « *Gazetteers are lists of objects of concrete type represented by strings that can be matched in the text. The well-known gazetteers from the existing IE tools are gazetteers for geographical location names, lists of organizations or people* » et emploie aussi des expressions régulières. Ceci afin d'analyser les conversations emails et d'en extraire des informations.

Plus proche de nos travaux, on peut citer une analyse très détaillée (Soares, 2013) sur les Knowledge Intensive Process (KIP) et comment en garder une trace. A l'aide d'outils de NLP, des réseaux sociaux des emails, et du story mining (Goncalves 2009), des mindmaps des conversations emails sont générés.

Mais si ces méthodes sont praticables sur des emails d'un individu, elles sont insuffisantes dans le cas d'email d'un groupe de personne ou d'un projet comme dans notre étude (même avec des sens identiques, deux emails différents n'auront peut-être pas de mot clef en commun).

Cette remarque trouve un écho dans les travaux de Weerkamp (2009) sur la recherche d'email, qui considère que les emails ne sont pas isolés, mais font partie d'un environnement en ligne plus large. Ce contexte, existant à différents niveaux, peut être incorporé dans le modèle de récupération. Aussi, il rajoute dans son système de recherche l'utilisation des fils de conversation, liste de diffusion, et les niveaux de contenu communautaire, en élargissant sa requête d'origine avec des termes de ces sources extérieures.

Cet aspect est très important dans notre approche centrée connaissance car les traces que nous cherchons n'existent et n'ont de sens qu'en contexte.

4.16 Extraction de Réseaux Sociaux

Enfin il y a des systèmes qui emploient l'analyse de réseau sociaux sous-jacent aux emails (Stolfo, 2006). Dans l'étude de (Laclavík, 2010) un réseau social multidimensionnel est construit à partir des emails (avec des liens sur les dates, les personnes, les sujets, les urls, les détails des contacts etc.). On crée un graphe bipartite éventuellement orienté (les nœuds représentent les adresses des participants, et les arêtes représentent le fait qu'il y ait un email de reçu/émis entre ces deux adresses). (Laclavík, 2011) étendra son approche avec des expressions régulières et des gazetteers.

La démarche réseaux sociaux est attractive car elle remet l'humain et les relations au centre du problème en ne considérant pas seulement des données textuelles, cependant elle est parfois difficile à mettre en œuvre sur de grand corpus, et représente mal le côté dynamique (temporel) des échanges.

4.17 Classification/Clustering

Certains travaux ont exploité le Machine Learning, par exemple avec succès les SVM (Support Vector Machine) (Drucker, 1999) pour classer les données textuelles. La recherche sur l'email s'est appliquée à créer des outils pour gérer les collections personnelles de courrier. Dredze (2006) a travaillé sur la gestion des activités emails en faisant du clustering. Peu après dans (Dredze, 2008) un « contexte humain » plus large est pris en compte, en classant des emails en fonction de topics (avec du LSA), et le vecteur caractéristique incorpore des parties du profil utilisateur (carnet de contact, manager ou supérieur, etc.)

Des chercheurs avaient abordé très tôt le problème de la classification automatique des emails (Helfman 1995), (Segal, 1999), (Cohen, 1996). Les premières techniques étaient axées sur un point de vue lexical/syntaxique et utilisaient des règles basées sur des mots/ngrammes clefs (souvent définis manuellement) et leur fréquence (en créant une estimation probabiliste de leur appartenance à une classe). Cselle (2006) a proposé un algorithme de clustering pour regrouper les emails et utilise une méthode heuristique pour labelliser les topics. Plus tard dans (Cselle, 2007), il tiendra compte

des informations concernant le destinataire, l'émetteur, thread, taille des messages, pièces jointes, pour créer les topics. D'autres tâches de classification concernant les emails utilisent le BOW (sur les sujet et corps) en les combinant avec les métadonnées (émetteurs, récepteurs, etc..) (Segal et Kephart, 1999) (Cohen, 2004) (Carvalho, 2007).

De manière similaire dans (Wajid, 2011) il y a regroupement de documents par type sémantique (en une ontologie locale) en utilisant Ontea (Laclavik, 2012) et la reconnaissance de patterns. La construction des labels de topics une fois les regroupements par cluster est faite automatiquement comme dans (Yang, 2010) où une combinaison du sujet et du corps permet de construire le VSM et un K-Means (MacQueen, 1967) pour les clusters. La labélisation des topics (des sujets) a posteriori (une fois les regroupements effectués) est une approche statistique qui nous semble moins prometteuse dans notre cas, car elle n'utilise pas la connaissance existante sur l'environnement projet.

Enfin il faut souligner les travaux de Feldman (1998) et Weng (2004) qui ont beaucoup influencé notre approche. Dans ces études, un ensemble de mots clef regroupé sous un vocable commun est utilisé pour des tâches de classification ou de détection. Feldman décrit le système KDT (Knowledge Discovery in Text), dans lequel les documents sont taggés (marqués) par des mots-clés, et la découverte de connaissances est réalisée par l'analyse des fréquences de cooccurrence de ceux-ci. Il applique ceci dans le cadre de KDD (Knowkedge Discovery In Database) mais indique que cette approche est très prometteuse pour les collections de textes non-structurés. Weng emploie une technique similaire pour classer des emails et recommander des réponses appropriées. On retrouve cette même technique dans Sakurai (2004). Ces dictionnaires de mots clefs trouvent leur fondement dans Sowa (2000).

Ces techniques utilisent les approches usuelles en NLP, BOW, etc. mais nous avons vu au chapitre 3 que les échanges par email peuvent être assimilés à une forme de conversation, aussi certains ont eu l'idée d'utiliser l'approche pragmatique et la théorie des actes de langages (AL).

4.18 Classification pragmatique

La pragmatique et les actes de langage (AL) ont été appliqué en informatique (Singh, 2006) et à l'email, en particulier pour classifier les emails en fonctions des intentions de l'émetteur.

Par exemple dans Cohen (2004) des classificateurs ont été mis en place basés sur des techniques de machines Learning pour de nombreux actes de langages. Chaque AL est décrit comme paire (verbe, nom), i.e. « Deliver, Commit, Request, Amend, ou Propose » et le nom. De manière analogue, Yelati (2011) effectue un taggage automatique des phrases dans des emails de support technique en fonction des intentions de l'auteur (sa grille d'analyse étant « Greet, Background, Goal, Concern, Query, Address ») et surtout il y ajoute de caractéristiques "sur-mesure"" dans la construction des vecteurs pour les SVM.

Dans (Carvalho, 2006) il est démontré que l'exploitation de l'information contextuelle dans les messages peut sensiblement améliorer la classification des actes de langage dans les emails (qu'il appelle « email acts »). Plus précisément, un prétraitement suivi d'une combinaison de séquence ngrammes se révèle très efficace pour cette problématique.

Goldstein et Sabin (2006) ont également travaillé sur des tâches similaires de classification des emails en employant des phrases clef construite manuellement, (pour détecter ce qu'ils ont appelé « email genre »). Des genres d'emails sont définis (memo, conversation, spam, doc officiel) et une sous classe des actes de langages identifiant l'intention majeure de l'émetteur avec son "genre" correspondant. Dans cette étude, les meilleurs résultats en classification sont obtenus en utilisant un lexique de verbe réduit et les caractéristiques de l'email.

Afin de découvrir les tâches (Khoussainov, 2005), un groupement des messages par date, émetteur, par similarité (BOW) est effectué puis utilisé afin de trouver comment les relations entre les messages d'une tâche peuvent aider à caractériser les actes de langages et en même temps symétriquement comment les actes de langages peuvent aider à découvrir les relations entre messages et les regrouper en tâches.

Des recherches ont également eu lieu sur la réalisation des actes de langage et les séquences d'AL (Jose 1988). Dans le travail de (Scerri, 2008), le but est de résumer les échanges d'email en actes de langages et de modéliser leur enchaînement pour les rattacher à des tâches et assister les usagers dans leurs travaux quotidiens. Scerri définissant l' « acte de langage email » comme un triplet (a,o,s) où a est une Action, o l'Objet de l'action et s le Sujet de l'action. De même dans (Carvalho, 2005), les e-mails sont étudiés pour voir s'ils contiennent des AL proches des requêtes ou des engagements, et il est indiqué que la corrélation entre les messages d'un même fil peut aider à cette détection.

Certains ont également étudié et formalisé des scenarios communs qui occurrent dans les conversations par email d'un point de vue des processus sémantiques (McDowell, 2004). Le travail sur la sémantique dans les communications par emails (Semanta (Scerri, 2007)) met en œuvre la théorie des actes de langage sur les e-mails. Cependant Semanta se concentre davantage sur la compréhension des AL que sur les informations contextuelle métier nécessaires à l'interopérabilité sémantique.

Enfin Lampert (2009) s'est penché sur le découpage de l'email en zones (par exemple, Salutation, Corps, Signature, « Disclaimer », Texte Cité et Copie de la réponse) avec un classificateur SVM (en prenant comme paramètres des informations, graphiques, lexicales, orthographiques), ceci dans le but d'optimiser dans un deuxième temps la reconnaissance des AL.

4.18.1 Annotation des AL

Une des difficultés inhérentes à ces tâches est le fait qu'il faille annoter manuellement les résultats (tagger les actes de langage pour l'apprentissage et/ou l'évaluation des performances) et que cette étape en elle-même, outre le fait d'être longue et fastidieuse, pose le problème du codage des AL et de l'accord entre les annotateurs (c'est subjectif et demande souvent une connaissance du domaine/contexte).

Il est important de remarquer avec l'étude de Lampert (2007) que même en cas d'annotation manuelle des AL, il n'est pas toujours simple d'obtenir un accord entre les annotateurs (dans cette analyse, un accord élevé sur les phrases comportant une

requête pour action (K^{14}=0.78) mais plutôt faible sur les engagements (K=0.54)). Aussi (Lampert (2009b) a poursuivi ces explorations des techniques de recueil via un complément Outlook permettant aux usagers de signaler eux-mêmes les AL. Plus tard, De Felice (2013) a présenté un algorithme d'annotation manuelle très précis et méthodique sur un corpus d'email professionnel.

4.18.2 Analyse du contexte des messages

Dans le travail original de Feng (2006) appliqué aux forums mais transposable aux emails dans son approche, il est souligné l'importance des AL pour analyser le sujet d'une conversation. Cependant la labellisation calculée des AL n'est pas suffisante pour comprendre le sujet central d'une conversation sans considérer d'autres caractéristiques comme des informations de l'auteur (la technique employée est particulière et relève d'un parcours de graphe assimilable au « random walk[15]» ou à l'algorithme PageRank (Page, 1999)).

En restant dans les analyses axées sur la pragmatique mais appliquée en dehors de l'email, on peut noter les travaux de Orkin (2001) sur l'analyse des dialogues d'un jeu pour les classifier en actes de langage qui prennent également en compte la situation physique dans l'univers simulé (contexte) et avec le codage DAMSL (Codage spécifique des actes de dialogue (Core, 1997)). De même (Ivanovic, 2005) emploie le même codage et une méthode statistique (naïve Bayes[16]) afin de modéliser et détecter les AL dans les messages instantanés (Site marchand MSN Online).

Un autre codage pour les actes de langage est le modèle VRM (Verbal Response Mode) (Stiles 1992) issu de la linguistique et de la psychologie. Le VRM prend en compte à la fois le sens littéral et pragmatique, mais aussi le point de vue de l'expérience de l'émetteur ou du récepteur. Ce codage a été employé par Lampert (2006).

---

[14] Le coefficient Kappa (K) en statistique mesure l'accord entre observateurs lors d'un codage qualitatif en catégories

[15] https://fr.wikipedia.org/wiki/Marche_al%C3%A9atoire

[16] https://fr.wikipedia.org/wiki/Classification_na%C3%AFve_bay%C3%A9sienne

Le choix des codages des AL est souvent abordé dans les travaux, si celui de Searle/Austin était employé au début, DAMSL a été davantage utilisé pour la détection des tâches, puis VRM, et enfin dans les derniers travaux, ce sont souvent des codages « sur mesure » adaptés des codages existants et aux problématiques abordées.

4.18.3 Détection des Requêtes

Dans notre étude, nous privilégions la recherche des actes spécifiques liés au contexte des messages. Ce qui peut réduire l'annotation manuelle des AL. Les interactions autour des projets sont entre autres liés à la résolution de problèmes et la coordination. Comme nous l'avons présenté chapitre 1, la résolution de problèmes peut être assimilée dans les interactions à des requêtes et des réponses. Nous allons examiner plus en détail un AL : la requête (détaillée au chapitre 3). La classification de la requête repose sur les idées fondatrices de Winograd (1986, 1987) pour prendre une perspective langage/action et identifier ces AL dans les emails.

La tâche de la classification automatiques en ce qui concerne la requête est compliquée car l'expression de la requête n'a pas une relation bijective avec sa réalisation linguistique.

Khosravi (1999) fut parmi les premiers à automatiser la classification des requêtes au niveau du messages en utilisant des règles mais bien spécifiques au domaine étudié. Or comme indiqué dans (Lampert, 2010) si analyser une forme surfacique particulière du langage est possible, il est reconnu que « enquêter sur une collection de forme qui représente par exemple un acte de langage entraine le problème d'établir *quelles* formes constituent cette collection » (Culpeper, 2008). Dans cet article, Lampert après une segmentation en zone, cherchera les « utterances » (phrase ou partie de phrase) mettant le récepteur en position d'obligation dans les emails. D'autres se sont intéressé à la classification des requêtes au niveau de « l'utterance », par exemple SmartMail (Corston-Oliver, 2004) visait à extraire des actions depuis les emails pours les inclure dans une « To Do List » (liste de tâche). Ce système employait un POS tagger, des aspects sémantiques avec des ngrammes.

Et bien souvent les requêtes sont associées aux actes de langage indirects (voir chapitre 3) et les emails ont montré qu'ils présentent une fréquence plus élevée d'AL indirects que les autres médias (Hassell, 1996)

Il faut citer le travail de Kalia (2013) visant à identifier automatiquement les tâches et la création d'engagement, la délégation, l'achèvement et l'annulation dans les conversations par courriel, basé sur des techniques de traitement du langage naturel et de l'apprentissage (SVM). Dans la même veine et avec des méthodes analogues (BOW, TFIDF, POS, SVM), les travaux de De Felice (2013) sur le traitement des emails dans les tests de TOEIC avec le codage VRM sont très proches. Dans ces cas, il s'agit d'apprentissage supervisé ce qui n'est pas sans poser de problèmes de mise en œuvre dans un cadre d'entreprise (annotateur).

4.19 Corpus Email

Enfin dans notre état de l'art, nous avons examiné les différents corpus ayant servi de support aux études. Il ressort que le corpus Enron (Diesner, 2005) est majoritairement utilisé, mais souvent sur de courts extraits alors qu'il comporte 250 000 emails. En second lieu, on trouve très souvent des corpus académiques issus des professeurs ou des étudiants, ou générés artificiellement d'exercices universitaires (PWCALO, Cohen 2004). Et enfin beaucoup plus rarement des corpus issus du monde l'entreprise.

On retiendra qu'il y a très rarement des corpus d'email professionnels (principalement pour des raisons de confidentialité) et que malheureusement les études sont souvent faites sur des échantillons de faible taille. (Sans doute en raison des annotations manuelles à faire mais ce qui peut questionner vis-à-vis de l'apprentissage automatique).

Nous avons fait figurer par ordre chronologique dans le tableau synthétique suivant les principaux articles sur lesquels sont basés nos travaux, et en grisé ceux pour lesquels, une idée particulière a trouvé un écho ou a été prolongée dans notre méthode présentée plus loin.

Récapitulatif tableau

Ref	Objectifs	Spécificité	Approches	Méthodologies	Emails	Corpus
(Feldman, 1998)	Classification de texte	Arbre de topics composés de mots clef	syntaxique	BOW, Mots Clefs	non	22173 textes de Reuters-
(Cohen, 2004)	Classification d'email	Ontologie des actes de langage	syntaxique, sémantique, pragmatique	BOW, TFIDF, SVM, Voted Perceptron, Decision Tree,	oui	4 datasets issus monde académique, nb messages (351,341,443) de Cspace et 222 de PWCALO
(Khoussainov, 2005)	Détection de tâches	prise en compte relation emails dans un fil	syntaxique, pragmatique	BOW, Actes de langage, SVM	oui	Corpus issu du mode professionnel 111 messages de 39 transactions d'affaire. (et PWCALO généré en exercice universitaire en jouant des rôles 6 personnes 4 jours)
(Diesner, 2005)	Exploration du Corpus Enron				oui	corpus ENRON email (Klimt and Yang, 2004) 250,000 email messages
(Ivanovic, 2005)	détection et modélisations des actes de langages	modélisation DAMSL des actes de dialogue	syntaxique, pragmatique	Actes de langage, Naive Bayes, ngrammes,	non	dialogue online shopping MSN, 9 dialogues
(Carvalho, 2005)	Classification Détection d'acte de langages	corrélation des actes de langage dans le même fil	syntaxique, pragmatique	Maximum entropy classifier	oui	249 + 137 messages de Cspace
(Feng, 2006)	Extraction d'information depuis message	Utilisation méthodes parcours de graphe, prise en compte d'information sur l'auteur	syntaxique, pragmatique	Actes de langages, VSM, TFIDF, algorithme a base de graphe (HITS)	oui	2414 messages d'un forum etudiant université

Année	Ref	Objectifs	Spécificité	Approches	Méthodologies	Emails	Corpus
2004	(Weng 2004) (Sakurai, 2004)	Classification de texte D'email	Dictionnaire de mots clefs pour les concepts	syntaxiques	BOW, TFIDF Machine Learning	non oui	612 FAQ de www.ntfaq.com 466, et 581 email du centre clientToshiba
2006	(Goldstein, 2006)	Analyse du genre	modélisation spécifique des actes de langage (email genre)	syntaxique, pragmatique	BOW, Actes de langage, SVM, POS, Random Forest	oui	280 messages au hasard de la collection personnelle de l'auteur
2006	(Carvalho, 2006)	Classification d'Actes de dialogue	Codage spécifique des Actes de Langage	syntaxique, pragmatique	Actes langage, Ngrammes,SVM	oui	1716 messages de Cspace (exercice étudiant)
2006	(Lampert, 2006)	Classification et détection actes de langage	codage VRM et mise en place de caractéristiques sur mesure	syntaxique, pragmatique	Acte de langage, SVM, réseaux bayesien, ngrammes, arbre dépendance	non	1368 utterances annotées de 14 dialogues du manuel de codage VRM de Stiles
2007	(Lampert, 2007)	Annotation Manuelle de Requests-for-Action (requête) and Commitments-to-Act (engagements)	Méthodologie d'annotation des actes de langages pour les requêtes et engagement	pragmatique	Manuel	oui	54 messages email contenant 310 phrases du corpus ENRON
2007	(Cselle, 2007)	Regroupement d'emails par topics	Labellisation des topics à la fin	syntaxique,	VSM,TFIDF,supervised clustering, SVM	oui	1586 messages pour les développement et 817 pour les tests issus d'une personne académique

Ref	Objectifs	Spécificité	Approches	Méthodologies	Emails	Corpus
(Dredze, 2008)	Résumé du message par mots clef		syntaxique, sémantique	VSM, LSA, LDA, Machine learning	oui	résumé mot clef : 7 usagers ENRON , prédiction reponse : 4 usagers 2391 messages, prédiction pièces jointe 15000 emails de 144 usagers
(Scerri, 2008)	Modélisation des actes de langages	prise en compte de la succession d'actes de langage dans des processus	syntaxique, sémantique, pragmatique	Actes de langages, Semanta, POS, Gate	oui	50 emails du corpus ENRON
(Lampert, 2009a)	Découpage en Zone	découpage en 9, puis 3 zones (salutation, corps et fermeture)	syntaxique	BOW, SVM	oui	400 messages email au hasard de ENRON
(Weerkamp, 2009)	Recherche de message	Augmentation de la requête par des données extérieures	syntaxique	Bayes	oui	30,299 fils (threads) de W3C list, 50 topics
(Lampert, 2009b)	Aide annotation des actes de langage	interface pour annotation des actes de langage	pragmatique	Pluggin Outlook, Actes de langage	oui	
(Lampert, 2010)	Détection de requêtes pour action	niveau message	syntaxique, pragmatique	BOW,ngrammes,SVM	oui	partie du corpus ENRON email (Klimt and Yang, 2004) 505 messages
(Laclavík, 2010)	Extraction de réseaux sociaux		syntaxique	Expression régulière, mots/phrases clefs	oui	partie du corpus ENRON email

Ref	Objectifs	Spécificité	Approches	Méthodologies	Emails	Corpus
(Yang, 2010)	détection de topics personnels	labellisation des topics des Clusters	syntaxique,	Vector Space Model, K means	oui	issu de 20 newsgroups, dataset de 399 emails
(Wajid, 2011)	Alignement sémantique de documents professionnels		syntaxique, sémantique	Expression régulière, mots/phrases clefs	oui	
(Laclavík, 2011)	Extraction graphe et réseaux sociaux	nomenclature à partir de chaine de mot clefs	syntaxique	Expression régulière, nomenclature	oui	Partie du corpus ENRON
(Wasiak, 2011)	Etude de la contribution au management	Données sur l'utilisation de l'email et des actes de langages en situation de projet	pragmatique	manuel	oui	800 emails d'un corpus du monde entreprise
(Yelati, 2011)	Classification par intention de l'auteur	Codage spécifique des actes de langage, caractéristiques construite manuellement pour le SVM	pragmatique	POS, Ngrammes, SVM, Classificateur a base de régle	oui	231 emails depuis le helpdesk des étudiants ou celui du support IT de l'université
(Orkin, 2011)	Classification d'Acte de dialogue	codage DAMSL et prise en compte contexte (situation physique)	syntaxique, pragmatique	unigram bigram, trigramme, SVM, HMM, Actes langage	non	5,200 logs du jeu

Ref	Objectifs	Spécificité	Approches	Méthodologies	Emails	Corpus
(Laclavík, 2012)	Extraction d'information	combinaison d'information dont les gazetteers (nomenclature à partir de chaine de mots sur mesure)	syntaxique, sémantique	Expression régulière, nomenclature, arbre sémantique, réseau sociaux	oui	50 emails en Espagnol d'un partenaire (Fedit)
(De Felice, 2012)	Identification d'acte de de langage	codage VRM	syntaxique, pragmatique	Actes de Langage, BOW, POS, Ngrammes, machine learning (maximum entropy classifier)	oui	1,163 textes de Test de TOEIC, 9284 utterance
(Soares, 2013)	Extraction d'information et de connaissance	Connaissance, KIP(Knowledge Intensive Process), Mindmap en sortie	syntaxique, sémantique	XML, VSM, NLTK,POS, Ngrammes, Réseaux Sociaux et Mindmap, Story mining	oui	730 emails d'une société media sur un KIP
(De Felice, 2013)	Annotation Manuelle d'acte de langage dans corpus professionnel	méthodologie très précise d'annotation des actes de langage	syntaxique, pragmatique	Manuel	oui	partie du corpus ENRON (263,100 mots/à peu près 20,700 actes de langage)
(Kalia, 2013)	Identification d'acte de de langage		syntaxique, pragmatique	Actes de langage, BOW, NER, Arbre de dépendance, base de règle, SVM	oui	4161 emails d'une employé d'ENRON avec 50 autres personnes

4.20 Principaux concepts retenus dans notre approche

4.20.1 Rappel

Pour rappel notre problématique concerne la traçabilité et la logique de conception dans le cadre de la mémoire de Projet et dans quelle mesure les emails professionnels peuvent y contribuer. De l'étude précédente des traitements textuels et en particulier sur l'email, nous allons dégager les axes principaux de notre méthode.

Replaçons notre problématique dans un contexte d'environnement professionnel pour mieux en cerner les enjeux :

Les données :

- Un corpus d'email issu d'un projet
- La description et les données du projet, les artefacts produits, documentations
- La description, les CV, compétences, rôle de l'équipe projet

Les objectifs d'entreprise

Exploiter au mieux ces données, notamment en termes de traçabilité afin de pouvoir capitaliser l'expérience, le savoir et la connaissance, assurer un éventuel rajout de fonctionnalités au projet, ainsi que la maintenance et le support/SAV aux utilisateurs.

Les enjeux de recherche :

Dans quelle mesure les emails associés à ces données peuvent-ils être utile et comment ?

Notre approche vis-à-vis du traitement du langage

Nous allons décrire dans le chapitre 5 notre méthode dénommée **KTR** (pour Knowledge Traces Retrieval, Recherche de Traces de Connaissances). Elle est globalement issue des traitements du langage, de la pragmatique et de l'ingénierie des connaissances. Comme toute démarche pluridisciplinaire, elle se situe à la croisée de différentes modélisations et techniques, et chaque domaine de compétence amènera son regard sur la situation et ses solutions.

En particulier le traitement du langage et la pragmatique vont nous permettre de :

- Traiter les données (qui sont textuelles)
- Définir le niveau de travail : pragmatique plutôt que simplement syntaxique
- Préciser l'axe central notre méthode : La recherche d'information (IR)

4.20.2 Choix et mise en œuvre :

L'objectif de notre travail est de rechercher des informations dans un ensemble de message email. Ces informations consistent en des éléments de connaissances sur le Projet, notamment sur la partie traçabilité du Design Rationale (Logique de Conception).

Retrouver des messages pertinents depuis une collection de document candidat (corpus) peut être envisagé comme un problème de classification (catégorisation) ou un problème de classement (ranking).

⇨ Nous allons traiter notre problématique comme un problème de classement dans lequel plusieurs propriétés de pertinence sont évaluées puis combinées.

La catégorisation entrainerait la problématique de la labélisation des classes qui, nous l'avons plus haut, est peu convaincante et difficile au niveau sémantique. En outre, les

cas d'utilisation de notre méthode impliquent un comportement dynamique, et non une catégorisation statique.

Plus clairement, on ne cherchera pas à catégoriser les messages dans l'absolu mais à les classer selon une demande de l'usager.

Il faut revenir brièvement sur le rôle et la place de l'email dans le management de projet. Pour ce faire, on s'appuiera sur une étude très structurante de (Wasiak, 2011) qui relate les manières dont l'email peut être utilisé dans la gestion de projet au-delà de la simple coordination de tâches.

La figure 4.6 suivante montre que le contexte projet et humain est primordial pour saisir la fonction et le contenu réel d'un email.

What topics the e-mail is about			*Why* the e-mail has been sent			*How* the e-mail content is expressed			
Product	Project	Company	Information	Management	Problem Solving	Socio Emotional		Task Related	
The artifact being designed	The project supporting the design process	The company supporting the project	Using e-mail to share information already existing	Using e-mail to direct, manage and organize	Using e-mail to discuss problems critically	Positive Reactions	Negative Reactions	Sharing	Requesting
Subclasses within each class									
Features Function Ergonomics Cost Performance Materials Manufacturing Specification	Risk Plans Team Quality Cost Time Manufacture Delivery Milestones Contracts Documents & Resources Administration	Stakeholders Economic issues Financial resources Human resources Physical resources Knowledge resources Tools and methods Practices & Procedures	Informing Requesting Clarifying	Managing, Confirming	Goal setting Constraining Developing solutions (solving) Evaluating Decision making Reflecting Debating Exploring	Shows solidarity Shows tension release Agrees	Shows antagonism Shows tension Disagrees	Gives opinion Gives suggestion Gives orientation	Asks for opinion Asks for suggestion Asks for orientation

Figure 4.6 : « Email and engineering Project management » extrait de (Wasiak, 2011)

Dans cette même étude, il fait mention aussi de la résolution de problème (« problem solving ») dont la traçabilité nous questionne et, là aussi, on note que l'email est souvent utilisé pour celui-ci.

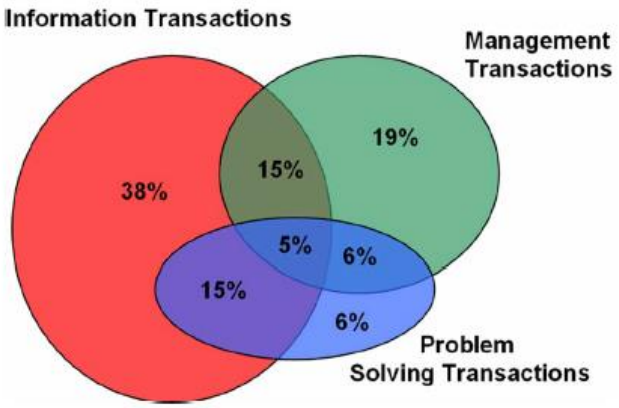


Figure 4.7 : « Frequency of use of the three transaction purposes of e-mail and the overlap between them ». d'après les travaux de (Wasiak, 2011)

Au final notre démarche s'articulera autour de 4 points :

1. Utiliser une démarche IR (recherche d'information) à partir d'une requête et l'appliquer au niveau message
2. Rechercher des messages contenant des informations relatives au projet
3. Rechercher des messages contenant certains actes de langage particuliers
4. Rechercher des messages dans lesquels les intervenants sont à même d'apporter des connaissances sur le sujet.

Pour ce faire à notre approche orientée connaissance et pragmatique accorde une place centrale au **contexte** : qu'il s'agisse du contexte projet dans son ensemble mais surtout par la prise en compte des intervenants et leurs relations et compétences (les acteurs du projet sont aussi importants que ce qu'ils écrivent).

Pour prendre en compte ces facteurs :

- Contexte projet : Nous allons introduire la notion de « **topics** » semblable aux « gazetteers » (Cunningham, 1995). Ce seront des dictionnaires de mots clef relatifs au projet. Ces Topics présentent également des analogies fortes avec la notion de « concept » de (Weng, 2004).

- La pragmatique est le bon niveau d'intervention pour le type d'information que nous cherchons aussi nous utiliserons des techniques de détection d'AL
- Contexte intervenants : Nous mettrons en place une modélisation des compétences et des relations (voir chapitre 5), celle-ci servira en partie pour la pragmatique

En pratique à l'échelle du traitement du langage :

De manière classique, nous emploierons le Vector Space Model (VSM) une modélisation BOW, et TFIDF sur les messages du corpus pour 2 aspects :

- tout d'abord retrouver les messages en rapport avec la requête et les classer par pertinence.
- ensuite trouver les messages en rapport avec les topics du projet et faire de même.

Sur les actes de langage et en particulier la Requête, de manière similaire à Lampert (2010), nous emploierons des techniques de Machine Learning (détaillées au chapitre 5) au niveau de la phrase. Comme dans (Cselle 2007), des informations concernant les émetteurs/destinataires seront employés et non pas seulement des aspects textuels.

Enfin comme nous l'avons indiqué plus haut la méthode KTR est dans le prolongement ou s'appuie sur les travaux de Feldman(1998), Sakurai (2004) et Weng(2004) pour notre modélisation des Topics, et sur ceux de Lampert(2009,2010), Carvalho et Cohen(2004,2006) et Kalia (2013) pour la partie pragmatique. Pour la partie intégration du contexte humain et projet, nous nous rapprochons de Feng (2006) et Weerkamp(2009)

Finalement nous pouvons situer notre méthode KTR par rapport aux approches existantes sur le schéma suivant l'usage de techniques NLP à plusieurs niveaux, l'orientation connaissance et la prise en compte de différents aspects du contexte:

Domaine	Ref	Utilisation de			Orientation		Contexte	
		Syntaxique	Sémantique	Pragmatique	Connaissance	Lexical	Projet	Humain
Information Retrieval (IR)	(Cohen, 2004) (Weerkamp, 2009)	***	**	parfois	-	***	*	parfois
Information Extraction (IE)	(Feng, 2006) (Laclavík, 2012) (Soares, 2013) (De Felice, 2012) (Kalia, 2013) (Dredze, 2008)	**		*	**	***	*	
Classification/ Catégorisation	(Orkin, 2011) (Carvalho, 2006) (Cohen, 2004) (Carvalho, 2005) (Lampert, 2006) (Yelati, 2011) (Cselle, 2007)	***	**		-	***		
Détection	(Lampert, 2010) (Khoussainov, 2005) (Yang, 2010) (Ivanovic, 2005)	***		**	-	**		parfois
KTR	(Rauscher, 2015)	**		**	**	**	***	**

Figure 4.8 : Approches NLP et problématique

Nous allons présenter en détail dans le chapitre suivant comment ces techniques de traitement du langage se combinent entre elles pour aider à la traçabilité des connaissances dans un projet.

Chapitre 5. METHODE KNOWLEDGE TRACES RETRIEVAL

5.1 Présentation

Dans ce chapitre, nous allons présenter en détails notre approche et la méthodologie associée. Notre système a pour nom KTR (Knowledge Traces Retrieval), l'objectif étant de repérer (trouver) dans un corpus d'email de projet d'entreprise des messages potentiellement porteurs de connaissances.

Nous allons définir les termes et le principe sous-jacent, puis nous détaillerons la méthode complète et l'algorithme support.

5.2 Principe général de KTR

Pour présenter de manière schématique l'hypothèse de notre approche, il est plus éclairant de partir d'un cas pratique. Un utilisateur a une question sur un projet abouti depuis quelques années. Cette question concerne la logique de conception : savoir pourquoi, comment ont été mise en œuvre telles propriétés de l'artefact produit résultant, quelles pistes ont été explorées, écartées, retenues et sur quels critères. Il a besoin de comprendre de ces éléments pour réaliser un nouveau projet similaire.

Il a sa disposition les données du projet, les profils intervenants et leur corpus d'email sur ce projet.

Une première solution est de faire une requête « full text » (plein texte) dans un logiciel d'email. Notre utilisateur va se retrouver confronté à des centaines de messages avec les mots clef correspondants à sa requête qu'il va falloir trier un à un afin de sélectionner les messages contenant vraiment des informations pertinentes. Notre hypothèse de solution comme nous l'avons évoqué plus haut se base sur le fait que la connaissance est produite suite à une interaction (Grundstein, 2000), que ce soit entre des acteurs ou entre un acteur et un ingénieur des connaissances. Dans notre approche, nous privilégions l'étude des interactions entre les acteurs à travers les e-mails.

Notre système KTR propose une analyse des e-mails afin d'explorer des sources de connaissances produites. Le but est de présenter les messages les plus pertinents à l'utilisateur en prenant en compte :

- Sa requête
- Le fait que les messages aient bien trait au projet
- Le fait que les messages fassent partie d'une classe d'acte de langage (AL)
- Le fait que les utilisateurs participant à ces messages aient en partie les compétences nécessaires à l'apport d'informations pertinentes.

Nous nous plaçons donc :

- Dans une approche IR (Information Retrieval) (Baeza-Yates et Ribeiro-Neto, 1999) : il s'agit d'assister l'utilisateur dans sa recherche d'information et non d'extraire automatiquement des informations ou des connaissances, ou encore de classer les messages par groupe.
- Dans une approche pragmatique : Nous allons chercher des actes de langages spécifiques en fonction des éléments que nous voulons tracer.
- Dans un contexte Projet : Nous allons employer les données et produits du projet, ainsi que des informations sur les intervenants et non pas seulement le contenu textuel des messages.

5.3 Niveau d'abstraction

En se plaçant à un niveau de modélisation, notre méthodologie est la suivante :

- Repérer les messages ayant trait au projet
- Repérer les actes de langages en rapport avec le type traçabilité souhaité
- Repérer les messages envoyés par des acteurs ayant les compétences requises pour fournir les éléments recherchés.

La figure 5.1 présente une vue globale des différentes étapes de KTR.

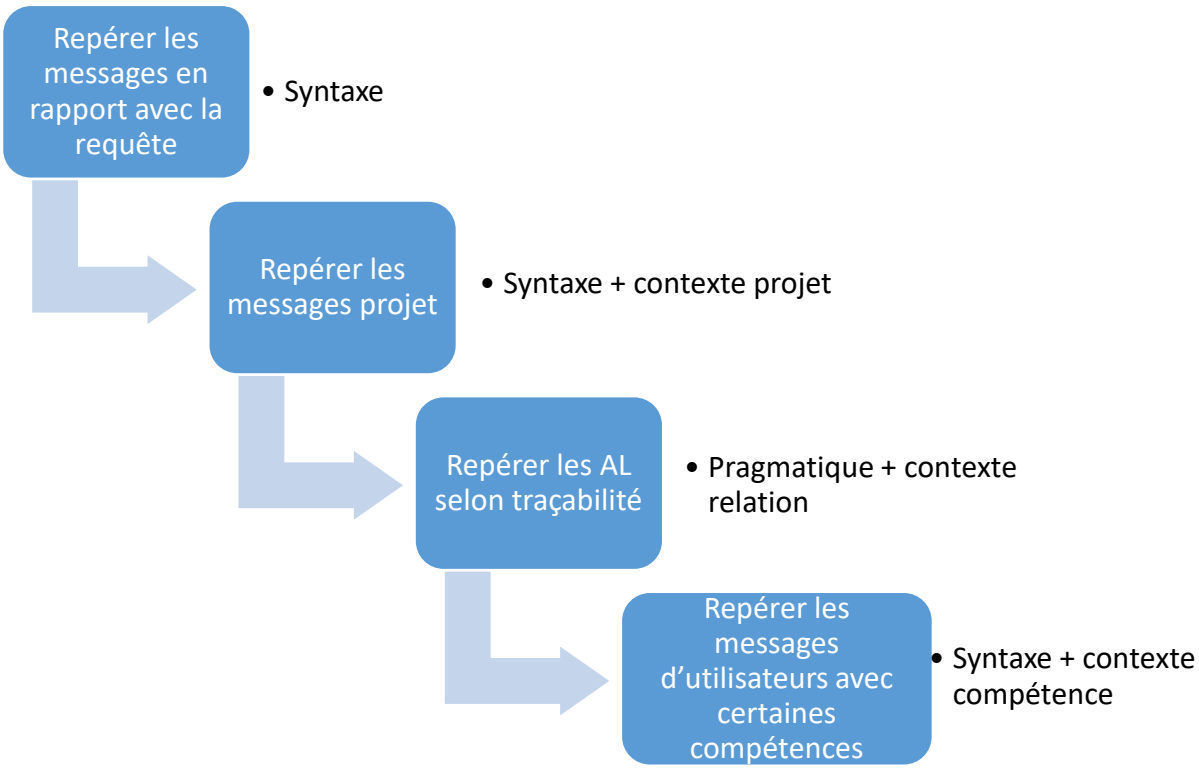


Figure 5.1 : Démarche globale de KTR

Il faut souligner qu'au final notre démarche finale n'est pas un filtrage incrémental (nous avons expérimenté cette voie dans (Rauscher 2015b), nous y reviendrons dans le chapitre 6). Il s'agit d'une combinaison de scores calculés dans ces sous modules afin de classer des messages par pertinence.

Dans cette étude, nous nous intéressons à la traçabilité de la logique de conception, et par extension à la résolution de problème (Newell, 1972) (MacLean, 1989), aussi

allons-nous nous focaliser sur un type particulier d'AL (Acte de Langage) : la requête afin de montrer la faisabilité de la démarche et la tester.

Cependant, notre méthode dans son acceptation la plus générale, n'est pas concernée par cette restriction. En effet, il serait envisageable de détecter d'autres classes d'AL (comme les promissifs, les expressifs, etc.. Searle (1969)) afin de tracer d'autres types de connaissances sur un projet (par exemple, les assertifs pour l'externalisation de la connaissance, les promissifs pour le suivi de la gestion de projet, etc.)

L'esprit général de notre approche est de prendre en compte un contexte large, en accord avec une ingénierie des connaissances pour l'action (Teulier, 2005). Dans cette vue, il n'est pas possible de faire abstraction des activités ou de l'organisation pour modéliser les connaissances. Teulier indique clairement en parlant des connaissances « C'est leur mise en œuvre, leur concrétisation dans une action pour atteindre un but qui nous intéresse car c'est bien l'activité et son caractère effectif que l'on cherche à assister. » (Figure 5.2)

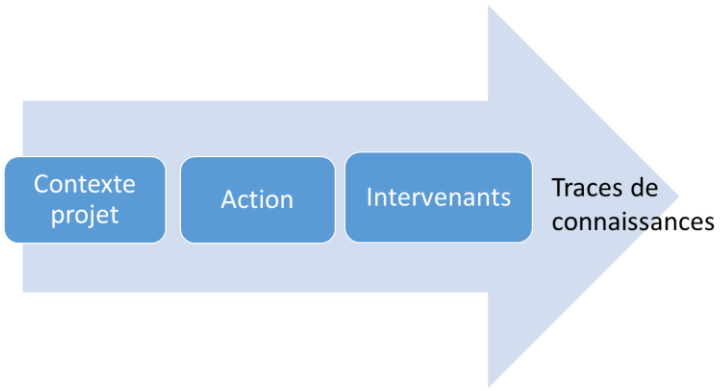


Figure 5.2 : Notre étude des traces des connaissances

Il y a eu beaucoup de travaux en Traitement du langage naturel ou NLP (nous l'avons vu au chapitre 4) sur la détection d'AL, sur la classification de texte, sur la recherche de texte ou d'expert mais aucune ne prend véritablement en compte les facettes multiples d'interaction avec la connaissance. Détecter un AL en lui-même ne suffit pas, pas plus que des mots clef ou des compétences particulières d'un usager. Pris individuellement, ces détections ou ces classifications sont des problèmes difficiles et présentent peu d'intérêt pour la traçabilité des connaissances. L'originalité de notre travail est d'essayer de combiner la détection de ces éléments dans un contexte projet

qui permette éventuellement de tracer la création, la transmission ou transformation de connaissance en la replaçant dans l'objectif global de l'organisation.

5.4 Description haut niveau

KTR est basé sur un calcul de score de pertinence sur les messages email d'un projet. Une partie du score sera issue de la similarité des messages avec les mots clefs de la requête utilisateur.

L'autre partie baptisée **KT_Score** proviendra de l'utilisation des thèmes « Topics » (voir chapitre 4), de la pragmatique et du contexte utilisateur et projet. Schématiquement on va donc évaluer :

- Un score_requete (message) : proximité du message avec les mots clef de la requête usager
- Un score_topic (message) : proximité du message avec les thèmes « topics » du projet
- Un score_al (message) : « probabilité » du message de contenir un AL
- Un score_compétence (message) : estimation des compétences des usagers sur les sujets du message
- Enfin un score_final (message) par combinaison des éléments précédents pour permettre de classer les messages par ordre de pertinence.

5.5 Focalisation sur notre étude

Comme indiqué au chapitre 2, notre travail porte sur la traçabilité de la logique de conception et de la résolution de problème dans les projets, afin de trouver les situations similaires appropriées proche de l'utilisation de la « signature » du TBR/CBR (Champin 2003). Dans ce dernier travail, il est considéré que des épisodes (partie d'une trace correspondant à une activité spécifique) liés à une tâche particulière

partagent généralement certaines caractéristiques communes. Une fois ces caractéristiques identifiées pour une tâche particulière elles sont considérés comme une signature de cette tâche. Leur instanciation dans la trace peut être interprétée comme une preuve de l'utilisateur qui effectue cette tâche pour la période correspondante. En effet, l'utilisation de la reconnaissance des signatures des connaissances pour l'extraction de traces de connaissance à partir de l'interaction homme-machine (surtout à travers les logs) permet de structurer la recherche des traces de connaissance selon un besoin particulier. Dans notre étude, la signature est basée sur trois éléments principaux : le thème, l'AL et les compétences de l'émetteur du message. Les traces des interactions ne sont pas des logs mais des e-mails.

Aussi en accord avec le chapitre 3, les type d'AL que nous allons prendre en compte seront les directifs et plus précisément les requêtes (directes et indirectes).

Nous avons vu qu'en phase de résolution de problème, ces AL sont très susceptibles d'être présents, et sans en être des marqueurs définitifs, ils font partie du faisceau d'indices pouvant indiquer cette phase. La phase de résolution de problème en conception sur le plan de la communication par email est marquée par des échanges avec des questions, des propositions, des décisions et des solutions par les membres de l'équipe projet. Ce sont ces messages là que nous voulons mettre en avant.

Cette démarche rejoint l'analyse de Teulier (2005) qui rappelle qu'en Ingénierie des connaissances, les méthodes de résolution de problème ne se construisent pas en «abstraction» totale de l'activité des acteurs et qu'au contraire la traçabilité et l'analyse de l'activité est « un moyen de voir les connaissances en action, de voir leur lien avec l'organisation, de les voir se traduire en choix, comportements, engagements ».

5.6 Approches des modules principaux KTR

Nous allons détailler les principaux éléments du score KTR et la manière de les aborder, à savoir les topics, la requête et la partie compétence.

5.6.1 Traces de connaissance

Comme indiqué plus haut, l'objectif de KTR est d'assister l'utilisateur dans la recherche de « trace de connaissance ». Dans la logique de ce que nous avons établi au chapitre 2 consacré à la traçabilité, nous définissons ici les *Traces de Connaissance* (Knowledge Traces ou KT) comme des messages emails comportant des informations significatives concernant les membres de l'équipe projet ayant un échange médiatisé sur une résolution de problème.

5.6.2 Les Topics

Le but est de décider dans quelle mesure un message contient des éléments significatifs pour le projet. Conformément à ce qui a été abordé dans le chapitre 4, notre approché présente des similitudes avec Sowa (2000). En effet nous n'allons pas regrouper les messages, puis déterminer par clustering des sujets, thèmes, « topics » a posteriori en nommant ces clusters.

Nous allons partir des données projet pour définir les topics. A la manière des « gazetters » (Cunningham, 1995), (Weng 2004), les Topics sont des dictionnaires de mots clef relatifs au projet.

Ils seront élaborés à partir des sources suivantes :

- Phasage du projet et documents de spécification, artefacts produits
- Ontologie du ou des domaines abordés
- Des experts sur le(s) domaine(s) du projet

Nous verrons plus loin dans la description formelle de KTR que certaines restrictions du NLP s'appliquent dans la construction de ces dictionnaires. Pour illustrer simplement cette étape, si nous décidons que le dictionnaire contient le topic « XML » défini par la liste de mots clefs « xml », « tag », « arbres », « xsd », « dtd », « balise », « structuration », « schéma », alors quand un message « contiendra » (aura une bonne similitude, voir plus loin la formalisation) une combinaison de ces mots clefs, nous saurons avec un certain score qu'il concerne le topic « XML ». Il faut noter que le nom des topics en lui-même est muet, il permet seulement une manipulation plus facile par les opérateurs humains.

5.6.3 La Requête

Comme indiqué ci-dessus, nous avons choisi d'abord de nous concentrer sur la résolution de problèmes car c'est le moment où les échanges entre les membres de l'équipe projets sont les susceptibles d'être porteurs de connaissances utiles à l'organisation. En particulier sur les « wicked problems » (Shum, 1997; Conklin, 1997, 2005), ou indéfinis. Conklin avait dégagé certaines caractéristiques de ces problèmes : ils sont nouveaux et uniques, n'ont pas de règle d'arrêt, la formulation de leur solution permet de les comprendre vraiment. Cependant la connaissance collaborative se manifeste plus naturellement dans le cadre de ce type de problème.

Par exemple dans le cadre du développement de logiciels informatique, lorsqu'une équipe tente de répondre aux spécifications et aux exigences du produit ou de corriger un bug, cela implique des tâches abstraites et cognitives. La première étant la qualification complète de la demande (le «but»), établir l'objectif et poser les problèmes pour l'atteindre. Les «opérations» se produiront dans les échanges suivants quand l'équipe va résoudre le problème (voir la partie suivante sur les compétences et les solutions). Dans notre étude, les équipes utilisent principalement la communication médiatisée par ordinateur, nous examinons donc attentivement les demandes/requêtes dans les discussions par courriel, puisque notre postulat est basé sur le principe qu'une question/requête peut correspondre à un problème à résoudre.

Nous avons vu au chapitre 3, que détecter des requêtes nécessite une interprétation de l'intention qui se cache derrière le langage employé. Aussi les techniques issues de la pragmatique linguistique seront employées. Cependant les actes illocutoires possèdent différentes modalités énonciatives : Il n'y a pas de relation bijective entre un AL et sa réalisation linguistique. Une requête est un énoncé qui permet de placer une obligation (de planifier/faire une action de répondre). Elle peut être directe (un ordre) ou, le plus souvent en environnement professionnel, indirecte (en questionnant la capacité, la volonté, etc ... de l'auditeur ou en donnant une suggestion).

Notre approche suit résolument une analyse du discours, nous travaillons sur des déclarations qui peuvent être interprétées correctement en tenant compte à la fois le contexte discursif (i.e. ce qui constitue le contenu du message e-mail) et le contexte de l'énoncé (c'est-à-dire la situation du dialogue).

Les travaux vus au chapitre 4 sur la détection de la requête nous ont montré les diverses approches. La plupart emploie des méthodes récentes issues de l'apprentissage automatique (Machine Learning). Elles utilisent un vecteur caractéristique sur l'utterance (énoncé), la phrase ou le message. Afin de construire ce vecteur, les auteurs fabriquent des caractéristiques numériques à partir de l'énoncé.

Il est intéressant d'en passer quelques-uns en revue afin de mieux comprendre nos choix.

Par exemple dans (Lampert, 2010) dans les composantes du vecteur sur un message destiné à l'apprentissage des SVM (détection de requête pour action), on trouve :

longueur du message en caractères et des mots;
le nombre et le pourcentage de mots capitalisés;
le nombre de caractères non alphanumériques;
si la ligne d'objet contient des marqueurs de réponses ou de transfert (par exemple Re :, Fw :);
la présence de noms d'expéditeur ou destinataire;
la présence de phrases qui commencent par un verbe modal (par exemple, pourrait, peut, devrait, serait);
la présence de phrases qui commencent par pronom interrogatif (qui, quoi, où, quand, pourquoi, qui, comment);
si le message contient des phrases qui se terminent par un point d'interrogation;
Présence unigrammes et bigrammes qui se produisent au moins trois fois à travers sur l'ensemble d'apprentissage

Figure 5.3 : Composantes vecteur caractéristique Lampert 2010

De même, on retrouve des éléments similaires chez (Goldstein 2006) dans la classification de type d'email, avec quelques manuelles caractéristiques comme la

présence des formules de politesse, d'excuse, de remerciement et moins d'éléments statistiques.

EMAIL CHARACTERISTICS FEATURES (EF)
Presence of Re:
Presence of Fwd:
Attachment signified in header info or by an insertion in text body
Fraction of interrogative sentences (sentences ending in '?'/total sent)
Fraction of "I" or "we" (count of words / total word count)
Fraction of "You" (count of words / total word count)
Attachment indicators such as "attached, here is, enclosed"
Apology indicators such as "sorry, apology, apologies"
Opinion indicators: "think, feel, believe, opinion, think, comment"
Politeness indicators such as "please"
Gratitude indicators such as "thank"
Action indicators such as "can you", "would you"
Commitment indicators such as "I can", "I will"
Information indicators such as "information", "info", "send"
Auto-reply indicators such as "out of the office", "away"
Email length

Figure 5.4 : D'après la classification et les travaux de Goldstein (2006).

Enfin dans les travaux de (De Felice, 2012), il est ajouté des données issues des POS (Part Of Speech voir chapitre 4) tagger, l'étiquetage morpho-syntaxique et la détection des entités nommées.

Feature	Possible values
Punctuation	; , . ! ? none
Length	Length of utterance
Subject type	Type: noun, pronoun, none
Subject item	Item: if pronoun, which one
Object type	As above
Object item	
Has modal	Yes/no
Modal is	Can, will, would, should, etc.
First word	Lexical item (excluding punctuation)
Last word	
Verb type	Infinitive, participle, etc.
Verb tag	VBD, VB, etc.
Sentence type	Declarative, question, embedded, etc.
Has wh-word	Who, what, when, why, where, how
Predicative adjective	Yes/no
Adjective is	Lexical item
Complex structures	I + modal + inf
	Please + imperative
	Be + adj + for me
	Etc.
Named entities	Place, time, name, organization, date, money
Unigrams	Small set of distinctive n-grams – cf. text
Bigrams	

Figure 5.5 : Vecteur caractéristique de De Felice 2012

Dans le système KTR, nous considérons que les marqueurs linguistiques seuls ne sont pas suffisants pour indiquer ce qui est réellement fait dans une situation de communication médiatisée par ordinateur. Nous allons reprendre les caractéristiques les plus simples des travaux antérieurs et y rajouter des éléments propres, selon nous, à mieux détecter les requêtes indirectes.

Nous avons choisi l'approche SVM (Support Vector Machine (Vapnik, 1995, 1998) ou Machine à vecteur de support, technique d'apprentissage supervisée qui généralise les classificateurs linéaires et qui obtient des résultats équivalent ou meilleur que les réseaux neuronaux) pour la détection de la requête car elle parait robuste d'après les travaux antérieurs. Cependant notre approche n'utilisera pas d'étiquetage morpho syntaxique, ni de statistique sur le nombre de mots, elle combinera au niveau phrase :

- Des éléments langagiers (n-grammes spécifiques, expressions, ponctuation, modaux)
- Mais aussi des éléments propres à la séquence du fil de discussion
- Des éléments concernant les relations de dépendance entre émetteur et récepteur(s).

Il faut bien saisir que dans notre méthode, nous sommes volontairement restés simples, nous ne voulons pas construire un détecteur de requête avec des performances au niveau de l'état de l'art mais que celui-ci s'imbrique dans un système plus global et qu'il participe à un faisceau de présomptions.

5.6.4 Les Compétences

Lorsque le système KTR a détecté une demande potentielle en ce qui concerne les thèmes des projets, nous aimerions suivre l'échange entre les membres de l'équipe afin de garder une trace des arguments, des questions connexes, les décisions et les solutions possibles au problème initial.

Notre hypothèse est alors que la connaissance de collaboration est plus susceptible de se manifester lorsque certains des intervenants qui échangent des messages ont les compétences nécessaires pour résoudre les problèmes actuels.

La définition des compétences dépend fortement de la discipline (sociologie, psychologie, gestion) comme indiqué dans (Harzallah, 2002; Vergnaud, 2004). Dans la perspective des ressources humaines, les compétences sont les connaissances mesurables ou observables, les qualifications, les capacités et les comportements nécessaires pour être performant dans son travail. On peut distinguer entre les compétences « douces » (managérial et interaction sociale), les compétences « dures » (fonctionnelles et techniques spécifiques à un domaine (Tripathi, 2014)). Dans notre modèle, nous allons mettre l'accent sur les compétences techniques et leurs relations avec les tâches qui doivent être accomplies pour le projet.

Plus spécifiquement, comme nous détaillerons cela dans la formalisation, nous allons lister les compétences techniques en rapport avec les topics du projet et évaluer les qualifications des intervenants sur celle-ci. En pratique cela se fera à partir d'une liste

de compétence, et avec les Curriculum Vitae « CV », description de poste, et rôle dans l'organisation du projet de chacun.

Ces grandes lignes étant définies nous allons présenter une description plus formelle de la méthode KTR.

5.7 Méthode KTR

Dans cette section, nous allons détailler chaque étape de notre méthode quel type d'information est utilisé, la construction des calculs de score, et comment le classement final pour la recherche est évalué.

5.7.1 Vue Générale

KTR suit une approche en deux étapes, d'abord une indexation des messages, puis un classement dynamique de ceux qui sont pertinents selon une requête de l'utilisateur. Afin de réaliser l'étape d'indexation, il est calculé des sous scores concernant les Topics, la Requête, la Solution selon des vecteurs caractéristiques. Cela fournit un score KT pour chaque message. Pour la partie du classement nous utilisons une combinaison linéaire du score KT avec le score de similitude basique entre la requête de l'utilisateur et le message.

La vue d'ensemble du système KTR est illustrée sur la figure 5.6.

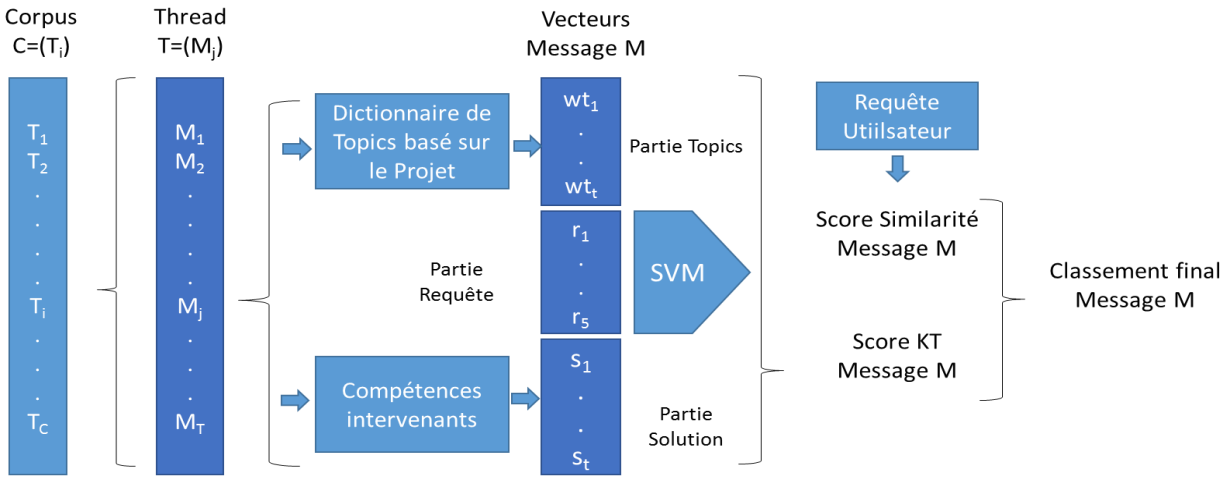


Figure 5.6 : le système KTR

Comme cas d'usage typique, l'utilisateur saisit une requête et le système KTR présente une liste de messages parmi les emails du corpus correspondant à la requête et ayant un score élevé de faire partie d'une séquence de résolution collaborative de problème entre les membres de l'équipe projet.

Pour mémoire, l'IR (Information Retrieval ou Recherche d'Information) peut être considérée comme un problème de classement ou de classification, dans cette étude, nous considérons comme un problème de classement (ranking). Nous calculons un score sur chaque message en fonction de la requête de l'utilisateur et les éléments que nous avons appelés KT (Traces de connaissance). Afin de décider si un message contient des KT, nous allons vérifier si:

- Le message traite des sujets du projet (Topics)
- Le fil de message contient au moins une Requête/demande (énoncé du problème)
- Les messages dans le même fil suivant la demande initiale contiennent des éléments de réponse ou une décision.

=> Il faut retenir que pour chaque étape, nous avons choisi les modalités de représentation les plus simples, (dictionnaire de topics, estimateur requête, solution/compétence). Ce n'est pas la complexité et la précision de chacune des représentations qui est visée, mais leur combinaison pour une prise en compte plus globale du contexte (projet et acteurs).

5.7.2 Définitions Communes

5.7.2.1 Corpus

Un corpus d'emails est un ensemble de messages ordonnés par heure d'arrivée et regroupés par fils (threads) (par exemple message initial avec des réponses).

Dans un thread T, comprenant les messages $(Mi)_{i \in T}$, chaque message M possède un émetteur E_M et des destinataires $R_M = (TO_M, CC_M)$ (respectivement destinataire direct (TO) ou en copie (CC)).

5.7.2.2 Prétraitement

Avant d'extraire les vecteurs caractéristiques pour le calcul des scores, une séquence d'étapes de prétraitements classiques en NLP a été appliquée à tous les emails avec quelques particularités propres au courrier électronique.

Les threads de messages ont été récupérés facilement via les fonctionnalités des serveurs comme Gmail ou Microsoft Exchange qui rajoute des identifiants thread dans les en-têtes (header). Cela permet un regroupement par fil qui dans les anciens logiciel d'email était assez fastidieux à reconstruire. Ensuite, à la manière de Carvalho (2006), nous avons supprimé les parties dupliquées des messages d'un même thread en cas de réponse ou de transfert, ainsi que les signatures et disclaimers (clauses de non responsabilité ou bannières de fin d'email). Puis nous avons anonymisé les corpus selon les demandes des fournisseurs. Enfin après suppression des espaces, et retour chariot supplémentaire, nous avons stocké les emails dans une base de données Microsoft SQL Server, en les découpant également en phrase.

La base de données comportait des tables correspondantes :

Threads
Messages
Phrases
Emetteurs
Destinataires_to
Destinaires_cc
Intervenants
Nom_pieces_jointes

Figure 5.7 : Extrait tables base de données

Ainsi que de nombreuses tables utiles à nos traitements (topics, requêtes, compétences, etc..).

Au final les parties restantes (en dehors des métas donnée de l'entête concernant la taille, date, expéditeurs, etc..) sont : l'objet de l'email, le corps assaini, et le nom des pièces jointes.

Récapitulatif prétraitement :

- Regroupement par thread
- Suppression des parties dupliquées, des signatures et disclaimers
- Anonymisation des intervenants
- Découpage en phrase et mise en base de données

5.7.3 Traitements

Comme indiqué précédemment, la méthode KTR se déroule en 2 étapes :

- une indexation avec calcul des 3 sous-scores correspondants aux Topics, Requête et Solution
- un classement final en combinant ces scores avec la requête utilisateur

Nous allons détailler ces étapes dans les sections suivantes.

5.7.3.1 Calcul score topics

Les topics sont directement issus des travaux de Sowa (2000), Blake (2001), repris par Feng (2004) où étaient définis la notion de « concept » comme « un groupe de termes capable d'exprimer le sens que l'auteur veut véhiculer ». Chaque « concept » peut être exprimé sous la forme Concept : {terme$_1$; terme$_2$; terme$_3$;…; terme$_n$}.

En utilisant les spécifications et phases du projet (ainsi que les sources citées plus haut), il est construit le dictionnaire de topics. Celui-ci est conçu de manière volontairement simple et prend la forme :

$L = (t_i)_{0 \leq i < t}$, où t est le nombre de topic, avec

Topic t_i : {motclef$_{i1}$, motclef$_{i2}$... motclef$_{ip}$}

Les mots clefs représentant les topics sont choisis de manière à ne pas trop se recouvrir de manière à garder les résultats significatifs. En pratique les topics se rapprochent d'un regroupement de taches correspondant aux étapes principales d'un projet.

Le but de cette étape est de savoir dans quelle mesure un message a trait aux divers topics. Pour cela, nous allons mesurer la similitude entre chaque message et chaque topic.

Pour ce faire, le contenu de tous les messages est représenté dans le VSM (Vector Space Model (Salton, 1975)) en utilisant le BOW (Bag of Word) avec le Porter Stemmer et une suppression des mots vides et avec une pondération TFIDF (term frequency, inverse document frequency), se conférer au chapitre 4 pour les définitions. Cela donne pour chaque message M

$M= (w_i)_{0\leq i<k}$ Dans lequel chaque terme est pondéré par son score *tfidf* , k étant la taille du vocabulaire

La même opération est effectuée pour représenter les topics. (On assimile un topic à un message constitué de ses mots clef).

Un « ranking » est alors calculé entre les messages et chaque topic en utilisant un algorithme basé sur la Cosine Similarity (voir plus bas implémentation pratique). Au final le résultat est une matrice T topics/messages où :

$T= (T_{ij})$ Avec (T_{ij}) représentant le poids du topics j dans le Message i $(0\leq T_{ij}\leq 1)$

Enfin le score pour la partie topic d'un message M_i sera donné par l'équation (1) suivante:

$$\text{Topic\_part}(M_i)=\frac{1}{t}\sum_{1\leq j\leq t} T_{ij} \qquad (1)$$

En pratique, pour l'implémentation, nous avons utilisé les outils Lucene (McCandless, 2010), et son dérivé Elastic Search (avec les parser et stemmer francais) pour calculer les scores et similitude. La formule de scoring de Lucene est un peu plus complexe (en particulier on peut accentuer (« booster ») certains termes lors de l'indexation ou de la recherche, c'est ce que nous faisons sur l'objet du mail).

$$\text{score}(q,d) = \text{coord}(q,d) \cdot \text{queryNorm}(q) \cdot \sum_{t \text{ in } q} \left(\text{tf}(\text{tin } d) \cdot \text{idf}(t)^2 \cdot \text{t.getBoost}() \cdot \text{norm}(t,d) \right)$$

Figure 5.8 : Scoring pratique de Lucene

On peut expliciter rapidement cette formule de la manière suivante : le score (similarité) entre un document d et une requête q sera calculé en utilisant les tf (fréquence des termes dans un document) et idf (inverse fréquence des termes dans tous les documents), coord (q, d) prend en compte le fait qu'un document contienne davantage de terme de la requête qu'un autre, querynorm(q) est un paramètre de normalisation des scores des requêtes, t.getBoost() permet de d'augmenter le poids de certains termes de la requête, et norm(t,d) permets d'accentuer certains champs et de normaliser à l'indexation.

Pour de plus amples détails se reporter à la page Lucene[17] de la fondation Apache.

[17] https://lucene.apache.org/core/3_5_0/api/core/org/apache/lucene/search/Similarity.html

5.7.3.2 Calcul score requête

La détection de la demande est un problème non trivial bien connu. Nous avons pris une approche simple similaire à Lampert (2010). Cependant, nous avons travaillé au niveau de la phrase et si une requête a été détectée dans l'une des phrases, le message a été classé comme demande.

Le découpage en phrase s'effectue de façon classique selon les signes de ponctuation et de paragraphe. Nous sommes bien conscients qu'une phrase peut contenir plusieurs actes de langage qu'un AL peut se matérialiser sur plusieurs phrases, mais la demande en tant qu'AL se retrouve fréquemment à l'échelle de la phrase.

Nous avons choisi des paramètres personnalisés en reprenant certains éléments de travaux antérieurs, comme les signes d'interrogation, certains n-grammes, (Cohen 2005,2006) (Goldstein, 2006), se référer à la section Requête plus haut dans ce chapitre pour le détail de ces travaux.

Un classificateur SVM a été mis en œuvre et implémenté en utilisant Azure ML (Plateforme de Machine Learning dans le Cloud Microsoft). Ce choix d'un apprentissage supervisé a été dicté par la littérature mais aussi par le fait qu'il puisse se généraliser d'un corpus à l'autre. Cependant il existe d'autres classificateurs similaires comme par exemple le Voted Perceptron (Freund, 1999), les Decision Trees (Schapire, 1999). On peut se référer à l'étude de Cohen (2004) pour des résultats comparatifs.

En prenant en entrée les phrases d'un message email, notre classificateur binaire prédit la présence ou l'absence d'énoncés de type requête dans celui-ci.

Pour ce faire, nous avons établi avec un linguiste un ensemble de caractéristiques personnalisées construites à partir de chaque phrase. Présence ou absence de: signe d'interrogation, bi grammes et trigrammes spécifiques sur la base pragmatique («vous devriez», « il faut », « pouvez-vous », etc.) et des mots-clés («question », «problème», « erreur », « qui », « quoi », « comment », etc.). Nous ajoutons également une partie temporelle en prenant en compte le fait qu'un signe de demande était déjà présent dans les messages précédents du même fil. En effet lorsque les utilisateurs sont confrontés à un problème, ils se posent généralement beaucoup de questions avant de

parvenir à un accord sur une solution. C'est un peu un phénomène auto catalytique en particulier sur les problèmes complexes donc sources de connaissances potentielles.

Un autre axe d'analyse est celui des relations entre les membres du projet. Les rôles dans l'organisation sont importants pour notre étude car ils pourraient aider à détecter les demandes indirectes. Par exemple, si un manager est en train d'écrire à un employé «Je voudrais (...)», il faut pour nous y voir un signe d'une demande implicite. Notre modèle prend les rôles en compte en utilisant la fonction officielle (hiérarchique) ou des relations d'affaires (client / sous-traitants). On retrouve des similarités avec les relations complémentaires dans Watzlawick (1979).

Nous construisons une matrice « relation » R en utilisant un graphe orienté pondéré représentant les liens hiérarchiques et les liens donneur ordre / fournisseurs. Les utilisateurs sont les sommets du graphe et les poids sur les bords apportent une mesure de "influence" l'utilisateur/utilisateur (valeurs réelles comprises entre 0 et 1). R_{ij} signifie la «capacité de commande" de l'utilisateur i à l'utilisateur j. Nous utilisons le max des récepteurs direct (TO) dans le cadre des messages.

R	M	E	F	EF
M	0	1	1	0,5
E	0	0	0,5	0,25
F	0	0	0	1
EF	0	0	0	0

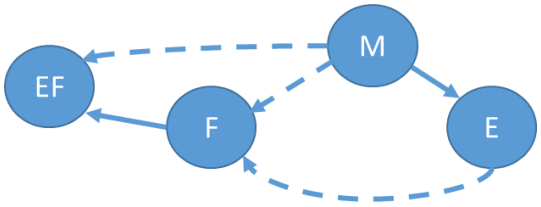


Figure 5.9 : Matrice « relation » et graphe

Exemple de Graphe et de matrice associée (M=Manager, E=Employé, F=Fournisseur, EF=Employé Fournisseur, les flèches pleines indiquent une dépendance hiérarchique, pointillées une dépendance contractuelle)

Au final, les caractéristiques que nous utilisons dans notre SVM détection de requête pour une phrase sont:

- Présence de signes verbaux (par exemple, pourrait, peut, devrait, pourrait, avez-vous, etc ..
- Présence de mots-clés spécifiques
- Présence marque d'interrogation
- Présence d'une requête détectée dans le même thread (message précédents)
- Le score d'influence des émetteurs / récepteurs

Il est important de noter que ce classificateur n'est pas spécifique au projet et une fois bien entrainés peuvent être utilisés dans d'autres projets.

Enfin, le score de la partie Requête est calculé comme dans l'équation (2), 1 si une phrase de requête a été trouvée dans le message M_i et Topic_part> 0 (nous rejetons les demandes n'étant pas liées aux thèmes du projet), 0,5 si le message faisait partie d'un fil où une demande antérieure a été trouvée, 0 ailleurs.

$$\text{Request\_part}(M_i) = \begin{cases} 1 \text{ si requête} \wedge \text{topic} \\ 0.5 \text{ post requête} \\ 0 \text{ pas de requête} \end{cases} \quad (2)$$

5.7.3.3 Calcul score « solution »

Nous recherchons des traces de connaissances (résolutions de problème) liées aux requêtes potentielles détectées à l'étape précédente. Pour chaque thread T, nous avons identifié des messages M_r où une demande est susceptible de se produire. Nous allons ensuite examiner tous les messages suivants dans le même thread c'est-à-dire $M_{s,T} = (M_i)_{(i>r),T}$ en prenant en compte les compétences des utilisateurs.

Tout d'abord, nous construisons une matrice CU représentant les compétences des utilisateurs, (en utilisant le curriculum vitae et la description de la fonction de leur rôle dans le projet). $CU = (CU_{ij})$ représentant le niveau de compétence utilisateur j dans la compétence i. A travers le modèle OntoProper, Sure (2000) utilise une base de profils

utilisateurs et crée des vecteurs numériques pour représenter les compétences de ceux-ci. Nous avons suivi une approche similaire à celle de Sure pour évaluer les compétences techniques (0 = ne connait pas, 0,25 = novice, 0,5 = moyenne, 0,75 = expérimenté, 1 = expert).

CU	Journaliste	Programmeur	Maquettiste
XML	0	1	0
BDD	0	0,75	0
Mise en page	0	0	1
HTML	0,5	0,5	0

Figure 5.10 : Exemple de matrice CU

Ensuite, nous avons construit une matrice CT représentant les compétences nécessaires pour être efficace sur un Topic (ses tâches associées) pour le projet, CT = (CT_{ij}) représentant l'importance de la compétence i sur le Topic j. Cette matrice est construite avec des experts dans chacun des sujets, encore une fois avec une pondération discrète (allant de 0 = compétence inutile pour les tâches, 0,25, 0,5, 0,75, 1 = compétence vitale).

CT	Import	Workflow	Magazine
XML	0,75	0,5	0,25
BDD	1	0,5	0
Mise en page	0	0	1
HTML	0	0,75	0

Figure 5.11 : Exemple de matrice CT

Nous construisons construit alors la matrice UT avec (UT_{ij}) représentant une estimation approximative des compétences de l'utilisateur i en ce qui concerne le Topic j.

$$UT = {}^{t}CU.CT$$

La matrice UT est normalisée en utilisant la norme de Frobenius[18] afin de ne pas trop dépendre du nombre de compétences ou d'utilisateurs.

Pour mémoire la norme de Frobenius (ou de Hilbert-Schmidt) pour une matrice A est

$$\|A\|_F = \left(\sum_{i,j} A_{ij} \right)^{1/2}$$

Enfin nous calculons sur chaque message $M_i \in M_{s,T}$, son score de « solution » (3):

$$\text{Solution\_part}(M_i) = \frac{1}{t} \sum_{1 \le j \le t} (UT_{ij}.T_{ij}) \quad (3)$$

Nous traitons des messages pouvant contenir des solutions potentielles concernant la résolution du problème soulevé par la requête de M_r, ce sous score est là pour prendre en compte le fait que l'émetteur dispose des compétences nécessaires pour apporter de nouvelles connaissances sur les Topics en cours.

5.7.4 KT Score

[18] https://en.wikipedia.org/wiki/Matrix_norm#Frobenius_norm

Le KT_Score (4) est calculé à partir des sous scores précédents sur chacun des messages M_i.

$$KT\_Score\ (Mi)= \begin{cases} Topic\_part(M_i)+ \\ Request\_part(M_i)+ \\ Solution\_part(M_i) \end{cases} \quad (4)$$

Ce score évalue la pertinence du message M_i à faire partie d'une trace de résolution de problème sur le projet.

5.7.5 Classement Final

Afin de procéder au classement final pour la phase de recherche basé sur la requête utilisateur Q, nous utilisons également la mesure de similarité basique (Cosine Similarity) dans l'espace VSM avec la pondération *tfidf* (avec l'implémentation de Lucène que nous avons évoqué plus haut).

$$sim(M_i,\mathrm{Q}) = \frac{\overrightarrow{M_i}.\overrightarrow{\mathrm{Q}}}{\|\overrightarrow{M_i}\|.\|\overrightarrow{\mathrm{Q}}\|}$$

Ceci afin de prendre en compte les termes spécifiques de la requête utilisateur.

Finalement le classement global r d'un message M_i est une combinaison linéaire calculée de la manière suivante:

$$r(M_i)= \mu Score\_KT_N(Mi)+(1-\mu)\mathrm{sim}(M_i,\mathrm{Q}) \quad (5)$$

Où μ ($o \leq \mu \leq 1 \in \mathbb{R}$) est un paramètre de combinaison (le KT_Score_N est normalisé). Avec $\mu=0$, nous revenons sur une similarité *tfdidf* classique comme dans le moteur de recherche de Lucène (Hatcher, 2004)

5.8 Algorithme

L'algorithme complet pour l'indexation et le classement de recherche est décrit dans la figure 5.12 page suivante.

Le classement global nous donne un score de pertinence pour chaque message. Nous affichons les messages par score décroissant mais pour une meilleure compréhension de l'utilisateur, nous les gardons groupés par fil de conversations (threads).

Entrées: Données Projet Data, Corpus, Compétences Equipe

Sortie: KT_Score par message

1 Indexation:

2- Préparation Dictionnaire de Topic L depuis les Données Projets

3- Préparation Matrice Topic (section 5.7.3.1)

4- Préparation matrices CU,CT and UT (section 5.7.3.2)

5- Préparation matrice dépendance Rôles R (section 5.7.3.3)

6- Apprentissage SVM sur les phrases de requête

7 Pour chaque thread T $\in$ Corpus

8 Pour chaque message M$\in$T

9 Calcul de Topic_part(M) (Equation (1))

10 Calcul de Request_part(M) (Equation (2))

11 Calcul de Solution_part(M) (Equation (3))

12 Calcul de KT_Score(M) (Equation (4))

13 fin

14 fin

15 Recherche:

Entrées: Requête Utilisateur Q, KT_Score, Corpus

Sortie: Classement global et messages

16 Pour chaque message M$\in$Corpus

17 Calcul de sim(Q,M) (section 5.7.5)

18 Calcul de r (M) (Equation (5))

19 fin

20 Affichage des messages par score r décroissant et regroupé par fil (thread).

Figure 5.12 : Algorithme KTR

Chapitre 6. APPLICATION

6.1 Présentation

Dans ce dernier chapitre, nous allons appliquer la méthode KTR à un corpus d'email. Afin d'obtenir des résultats significatifs, nous avons pu obtenir un corpus issu d'un projet réel du monde l'entreprise. En premier lieu, nous présenterons le corpus, le projet, et les intervenants. Ensuite nous ferons une analyse plus détaillée des échanges et des caractéristiques du corpus. Enfin nous préparerons les éléments nécessaires aux modules de la méthode KTR et nous appliquerons celle-ci. Puis nous étudierons les résultats, les performances et nous en tirerons des conclusions.

Le Projet et son corpus

Dans le cadre de notre travail, il est impératif de choisir une étude de cas respectant les critères suivant :

- Une organisation et une structure de projet encourageant l'utilisation et l'archivage des e-mails.
- Une organisation de projet répartie géographiquement et avec des modules très interdépendants afin que le courriel soit nécessaire.
- Un projet achevé ayant nécessité une ingénierie de conception interdisciplinaire mêlant plusieurs domaines variés.

6.2 Données du projet :

Un groupe d'édition de magazines et de livres avait en 2009 un projet de conception de logiciel qui a duré 2 ans. Une SSII a été engagée à distance pour créer un outil de workflow pour les journalistes et les avocats. Presque toutes les communications au cours des spécifications, implémentations, tests et livraisons ont été effectuées par e-mail. Le corpus a été collecté 2 ans après la fin du projet. Quelques années après l'utilisation de la solution, le serveur est tombé en panne, un certain nombre de documents ont été perdus. Le groupe d'édition a fait appel à la SSII pour résoudre le problème. Les documents contractuels ne sont pas suffisants pour comprendre les raisons de choix développés. La SSII a regroupé l'ensemble de documents contractuels dont elle dispose ainsi que le corpus des échanges (par e-mails) entre les acteurs. Notre approche KTR, comme nous l'avons précisé, se base sur la relation entre les échanges et le contexte afin d'isoler de traces de connaissances utiles dans la compréhension et la construction sémantique de la résolution de problèmes.

6.3 Le projet

6.3.1 Présentation

Le projet EDA (Les Editions de l'Argus de l'Assurance) est issu d'une demande du groupe INFOPRO Digital (groupe de Presse Professionnel qui regroupe environ 2000 personnes, des magazines, des logiciels et des salons).

Il s'agissait pour une Unité nommée Les Editions de l'Argus de l'Assurance de se doter d'un logiciel permettant d'informatiser la production de leurs codes et ouvrages et la publication cross media (web, papier).

Les Editions de l'Argus de l'Assurance produisent des Codes commentés (Code de l'Assurance, de la Route et de la Mutualité) ainsi que des ouvrages juridiques (par exemple l'assurance en Construction).

Ces codes sont constitués à partir des codes issus de l'assemblée nationale et mis à jour quotidiennement. Les Editions de l'Argus de l'Assurance y rajoutent des éléments éditoriaux (bibliographie, commentaire, texte attachés et annexes) à l'aide d'auteurs (avocats, cabinet, juriste, etc.) renommés dans le domaine.

6.3.2 Objectif

INFOPRO Digital s'est rapproché de la société SSII1 en 2009 car celle-ci avait été l'architecte en 2002 d'un Workflow pour l'ensemble des journalistes et maquettistes du groupe et toujours en production.

Dans l'esprit le projet EDA présentait des similitudes avec ce workflow existant. Il s'agissait de permettre à des journalistes et auteurs, de travailler via MS Word dans du contenu très structuré (textes juridiques en XML) et d'y rajouter des parties. (Techniquement cela revenait à une grande combinaison d'arbres XML). Une interface devait permettre la circulation des documents entre les journalistes et enfin plusieurs exports (papier, auteur et web) seraient finalement mis en place.

Le tout devant devenir l'outil de production pour Les Editions de l'Argus de l'Assurance.

6.3.3 Résultats

L'outil a été mis en début de production fin 2009, puis a pris sa vitesse de croisière (spécifications complètes) en juin 2010.

En 2016, il est toujours en place et assure la production quotidienne, de nombreux codes et ouvrages sont venus se rajouter et une acquisition prochaine du Groupe INFOPRO devrait voir son périmètre doubler. Il n'existe pas de solution similaire dans ce domaine.

6.3.4 Contraintes

Ce projet devait permettre la sortie d'un code papier en septembre 2009 au plus tard.

Puis la mise en ligne (sur le site internet du groupe) du même code, dans un délai le plus court possible. Des enrichissements (liens inter et intra documents) devaient valoriser les codes et ouvrages.

Les usagers devaient pouvoir utiliser les outils qu'ils connaissaient et maitrisaient dans une large mesure (Microsoft Word, Adobe Indesign)

Les processus devaient être automatisés au maximum sans redondance d'informations ni de manipulations. Le contenu issu de Légifrance ne devait jamais être touché car ce sont des textes de lois non modifiables.

Enfin l'ensemble du système devaient être extensible à un nombre quelconque de code et d'usagers.

Ce projet ne devait pas dépasser une enveloppe budgétaire de 120k€ environ.

6.3.5 Acteurs

Les sociétés ayant participé à ce projet :

INFOPRO (DSI, Les Edition de l'Argus, le Web) : Maitre d'ouvrage et client

SSII1: Maitre d'œuvre du projet, SSII

SSII2 (sous-traitant société SSII1, SSII en Edition)

LEGIFRANCE (Fournisseur officiel des Codes et de la Jurisprudence Française)

Remarque : Les noms des SSII ayant participé ont été anonymisés à la demande des fournisseurs du corpus.

6.4 Organisation du projet

6.4.1 Le Planning

Initialement, le projet a été planifié sur une période d'un an selon cinq phases essentielles (voir planning figure 6.1) : Extraction de documents en format XML, conception d'une Base de données, Définition du workflow, spécification d'une maquette et développement d'une solution Web.

PHASE	Sous-Phase	Date début	Date Fin	Ressource
Import XML		02/02/2009	02/03/2009	
	Récupération et spécification			Infopro
	Nettoyage Code et Ouvrage			SSII2
	intégration			SSII1
BDD		16/02/2009	02/03/2009	
	Conception			SSII1
	Implémentation			SSII1
Workflow		16/02/2009	23/03/2009	
	Spécification			Infopro
	Conception			SSII1
	Implémentation			SSII1
Maquette		02/03/2009	30/03/2009	
	Spécification			Infopro
	gabarit code			Infopro
	gabarit ouvrage			Infopro

	Implémentation			SSII1
	Service			SSII1
Web		01/04/2009	30/04/2009	
	Spécification			Infopro
	Implémentation			SSII1
	Lien Word			SSII1
Gestion projet	Transverse	01/01/2009	01/06/2009	Infopro
	Gestion éditoriale	01/01/2009	01/06/2009	Infopro

Figure 6.1 : Planning initial au 01/2009

En réalité, le phasage du Projet a subi des décalages et nous avons récupéré depuis des documents de gestion de projet, les éléments du phasage réels (Figure 6.2)

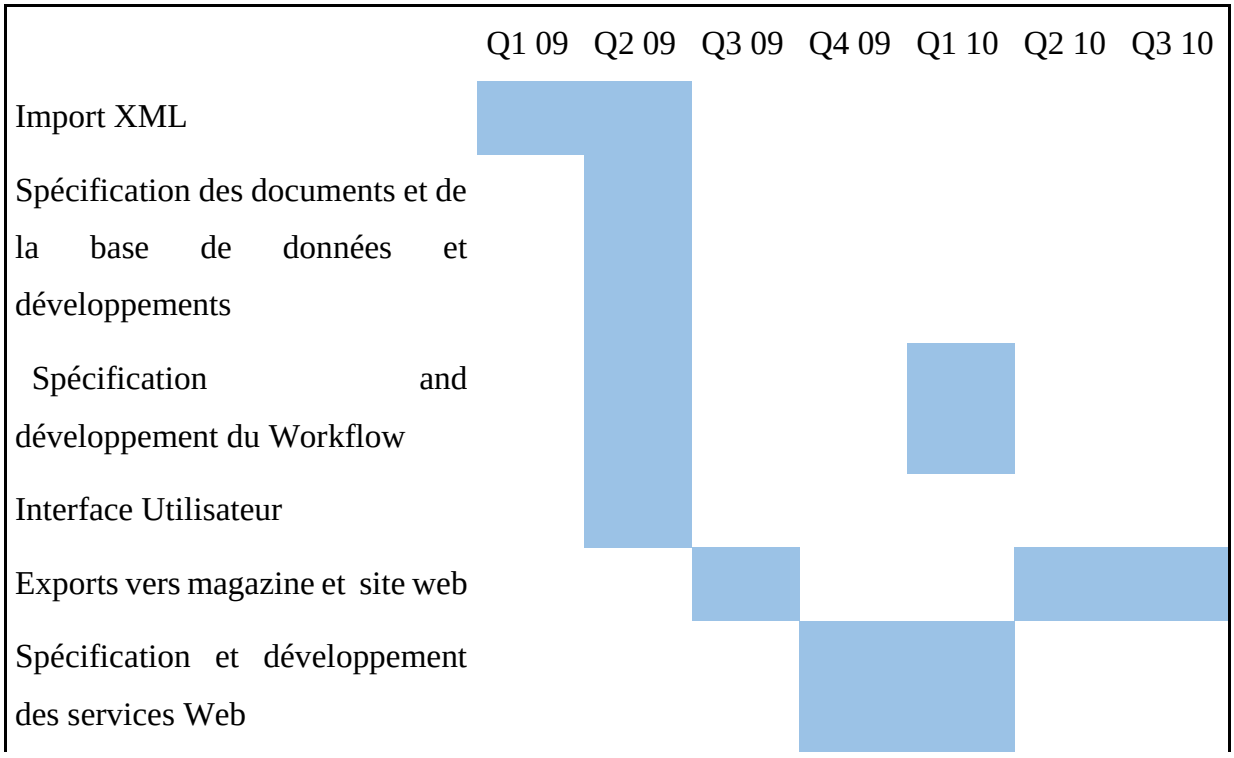


Figure 6.2 : Planning réel du projet

6.4.2 Présentation des utilisateurs

KTR est basé en partie sur l'identification de rôles et des fonctions des acteurs impliqués dans les interactions. Il est donc important de décrire les différents acteurs du projet que ce soit les futurs utilisateurs de la solution que les fournisseurs de celle-ci.

1.4.2.1 Fonctions des acteurs

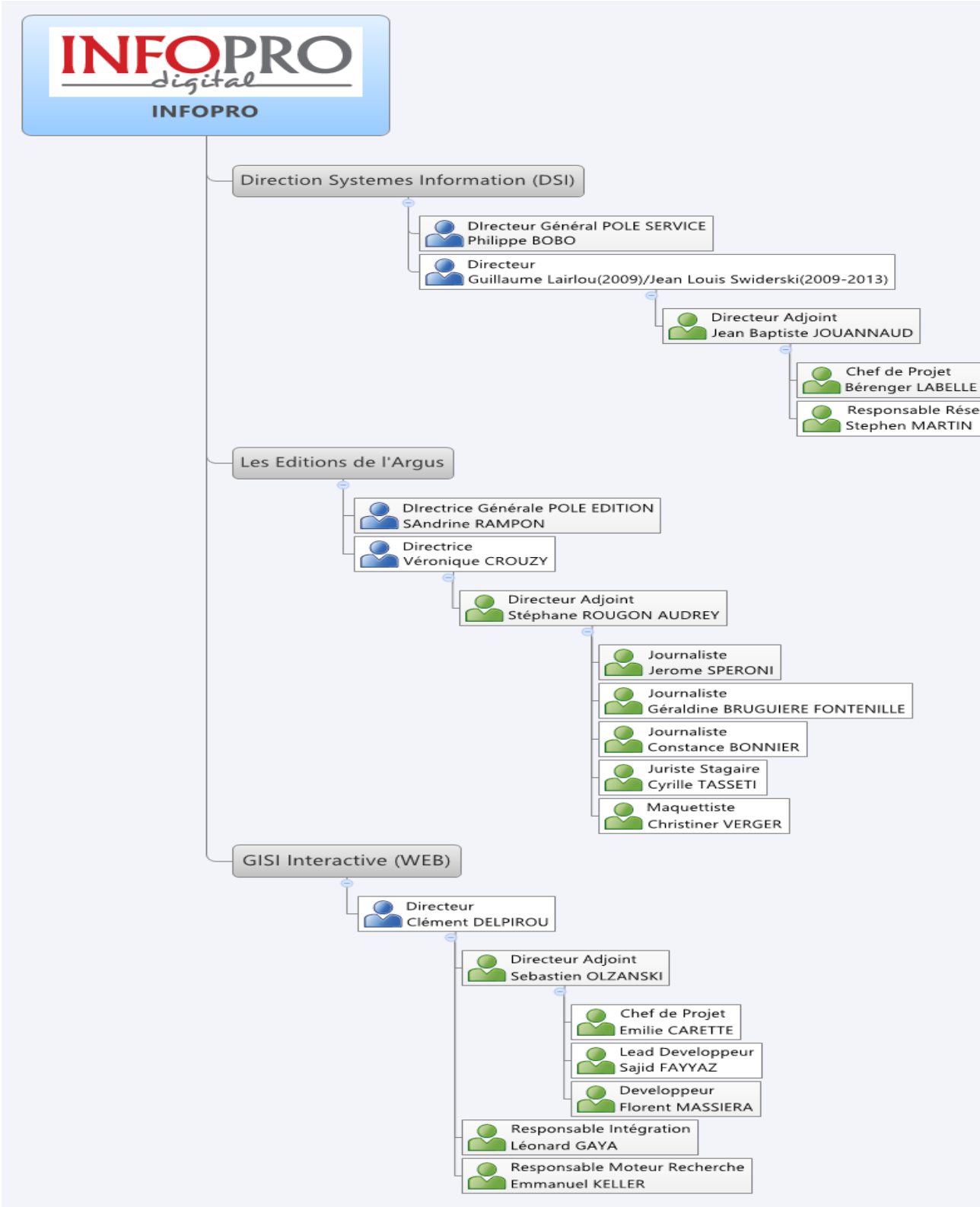


Figure 6.3 : Acteurs Infopro

SSII1

Directeur, Architecte Technique : FX

Développeur C# : MHU

SSII2

Commerciale : CRO

Développeur XML : RTO

LEGIFRANCE

Responsable : Pierre LAREDE (PLA)

1.4.2.2. Rôles dans le projet

Jean Baptiste JOUANNAUD (JBJ) chez INFOPRO a géré le projet en tant que MOA, a fait le fil conducteur entre les intervenants et pris les décisions plannings et budgétaire. C'était aussi un informaticien qui connaissait les contraintes techniques.

Stéphane ROUGON ANDREY (**SRA**) a géré le projet en tant que Client final et MOA Opératif, c'est un juriste de formation, journaliste, et ouverts aux aspects technologiques. Il s'est révélé un atout précieux sur ce projet avec une capacité de travail très importante. Au point que sa carrière a pris une nouvelle orientation en 2012 et qu'il est allé travailler comme directeur des projets PRESSE chez SSII2.

FX est Maitre d'œuvre de ce projet chez SSII1, il a conçu et développé toute l'architecture technique sous-jacente les bases de données, les flux XML,

l'orchestration, les web services et les interfaces utilisateur. Il a travaillé principalement à distance (via VPN) avec des réunions bi mensuelles en présentiel.

RTO est un sous-traitant d'SSII1, c'est un spécialiste du traitement XML. Il a été aussi embauché pour assurer une présence physique chez INFOPRO.

Jérôme SPERONI, Géraldine BRUGUIERE, Constance BONNET, Cyrille TASSETI sont des journalistes de l'Argus utilisateurs finaux de l'outil et sans profil technique.

Christine VERGER est la maquettiste des Editions de l'Argus et son avis est primordial puisqu'elle à la main sur le Code final/ (graphiste mac et a bien assimilé le XML)

Sajid FAYYA et les développeurs Web sont intervenus en fin de projet lors des exportations et sur des aspects uniquement techniques. Les relations entre eux et le reste de l'équipe ont eu lieu quasi exclusivement par email.

Pierre LAREDE était l'interlocuteur Légifrance, très compétent sur le côté juridique, la technique a été sous-traitée à la société ATOS. Les légistes lui ont fait d'ailleurs remonter des bugs sur les données Légifrance au cours du projet.

Philippe BOBO, Sandrine RAMPON, Clément DELPIROU sont des dirigeants principaux chez INFOPRO intervenus en rôle consultatif ou recette en début et fin de projet.

6.4.3 Tableau de rôles principaux

PHASE	Sous-Phase	Nom Code	Fonction dans l'entreprise	Rôle	Compétences
Import XML					
	Récupération et spécification	SRA	Directeur adjoint EDA	MOA Opérationnel	juridique, management
	Nettoyage Code et Ouvrage	RTO	Ingénieur	Développeur	XML
	intégration	FX	Architecte SI	Intégrateur	XML, BDD Développement,
BDD					
	Conception	FX	Architecte SI	Intégrateur	XML, BDD Développement,
	Implémentation	FX	Architecte SI	Intégrateur	XML, BDD Développement,
Workflow					
	Spécification	SRA	Directeur adjoint EDA	MOA Opérationnel	juridique, management
	Conception	FX	Architecte SI	Intégrateur	XML, BDD Développement,
	Implémentation	FX	Architecte SI	Intégrateur	XML, BDD Développement,

Maquette					
	Spécification	SRA	Directeur adjoint EDA	MOA Opérationnel	juridique, managemen
	gabarit code	CVE	Maquettiste		Maquette
	gabarit ouvrage	CVE	Maquettiste		Maquette
	Implémentation	FX	Architecte SI	Intégrateur	XML, BDD Développe
	Service	FX	Architecte SI	Intégrateur	XML, BDD Développe
Web					
	Spécification	SFA	Chef de projet Web	Lead Développeur	XML, Web
	Implémentation	FX	Architecte SI	Intégrateur	XML, BDD Développe
	Lien Word	FX	Architecte SI	Intégrateur	XML, BDD Développe
Gestion projet	Transverse	JBJ	Directeur Adjoint SI	MOA	Management
	Gestion éditoriale	BLA	Chef de Projet	MOA	Management, Maquet

Figure 6.4 : Tableau des rôles et fonctions

6.4.4 Versions et livrables

Initialement, il était prévu les livraisons suivantes :

- Version 0.5 pour intégration 16/03/2009
- Version 1.0 01/04/2009
- Version 1.1 avec export web le 01/06/2009

En réalité, l'historique des versions fut peu ou prou le suivant (toujours d'après les documents du gestionnaire de projet):

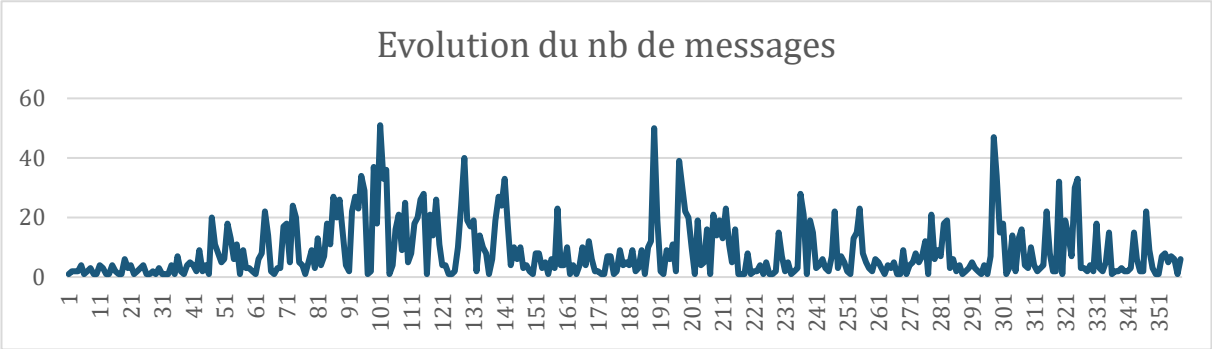
01/01-01/03/2009	**Analyses et premier développement**	**V 0.0a**
01/04/2009	changement complet de cadre (utilisation de legifrance) => quasi reboot du projet	V 0.1
01/05/2009	Première interface (vide) pour login/test	v 0.9
01/06/2009	première intégration des XML	v 0.95
01/08/2009	sortie d'un code test (sans word enrichi)	V1
01/12/2009	stabilisation interface et code	V 1.1
01/02/2010	ajout de la recherche full text	V 1.2
01/03/2010	web service dans word	V 1.3
01/06/2010	projet quasi fonctionnel comme prévu	
01/09/2010	Export web et ouvrage	v 1.4
01/01/2011	Projet en Production complète	v1.5

6.5 Analyse du corpus des échanges

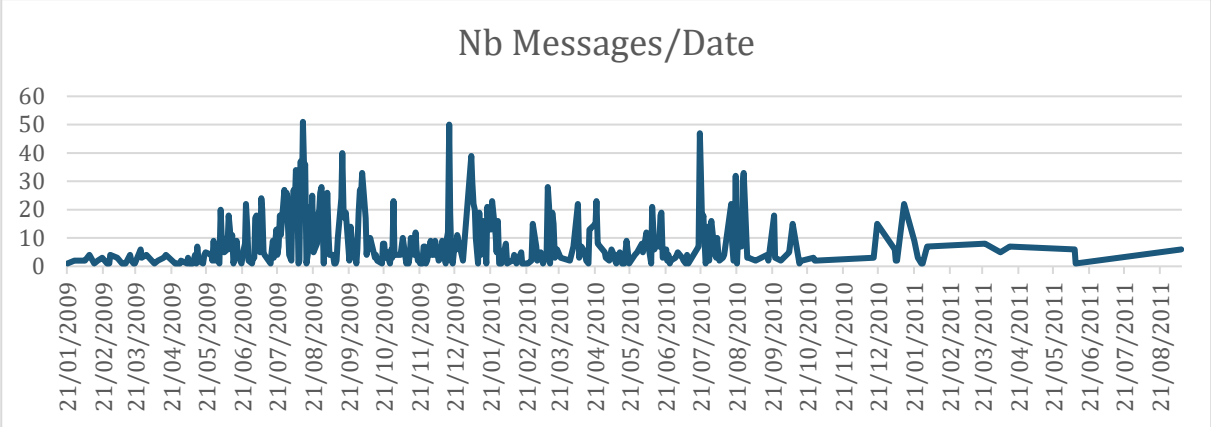
6.5.1 Eléments généraux :

Un Corpus des e-mails a été recueilli entre le 21/01/2009 et le 08/09/2011. Ce corpus représente 3080 messages / 14987 phrases dans environs 800 fils de conversations (threads) entre 30 acteurs du projet. L'équipe a été divisée entre les donneurs d'ordres et l'équipe de développement. Il faut souligner que le recueil du corpus s'est fait à partir de boites emails des intervenants selon le rangement propre à chacun (dossier relatif à ce projet). Il est donc aussi complet que possible mais nous ne pouvons être certains qu'il contient l'intégralité de tous les messages (certains ont pu être supprimés ou mal rangé par les intervenants). Parmi ceux-ci, différents rôles et compétences étaient présents. Comme le stipule KTR, il est important d'identifier les échanges liés à la résolution de problèmes. Pour cela, nous avons calculé des occurrences afin d'identifier les messages susceptibles de témoigner d'un échange riche entre les acteurs sur des problèmes significatifs du projet.

6.5.2 La répartition des messages en fonction du temps



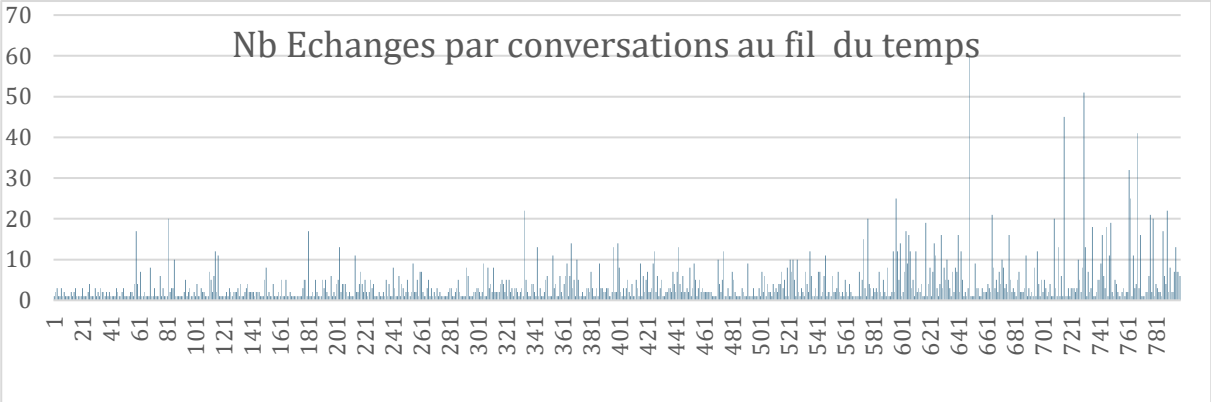
En temps incrémental relatif, on ne peut distinguer des pics et dégager ainsi plusieurs périodes d'activité.



En temps absolu, on peut distinguer les phases essentielles du projet qui ont donné lieu à un flux de messages plus important, puis le passage en production avec un plateau (voir le phasage réel plus haut). La table de fréquences ci-dessus montre 3 pics d'activité qui correspondent à des temps critiques du projet : la première et la seconde livraison et l'ajout de nouvelles fonctionnalités Nous avons identifié 10 intervenants dans ce projet qui comptabilisent plus de 80% des messages.

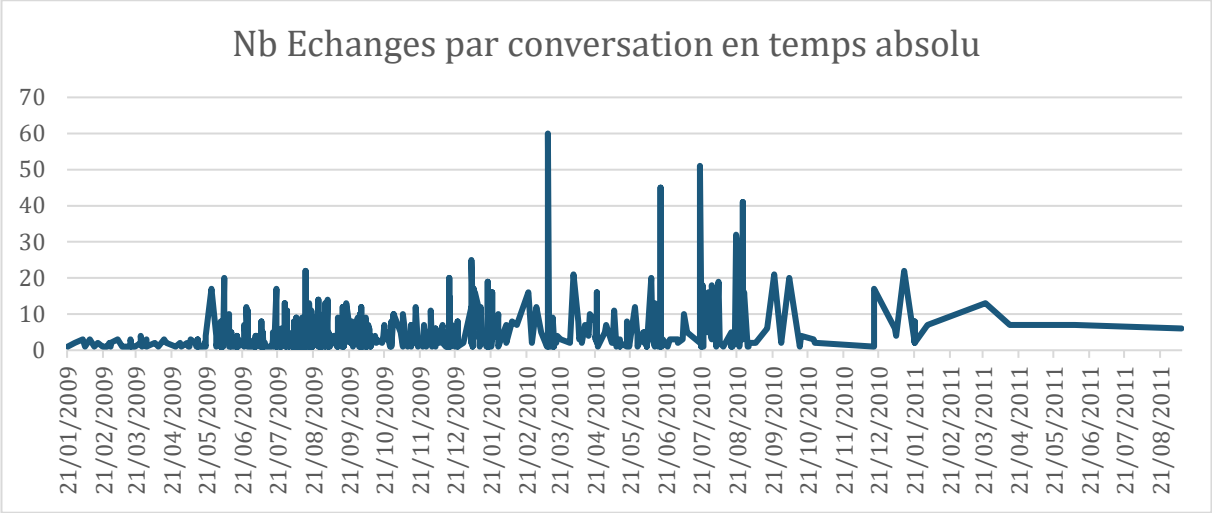
On constate une moyenne de 8.6 messages par jour.

6.5.3 Les fils de conversations (threads)

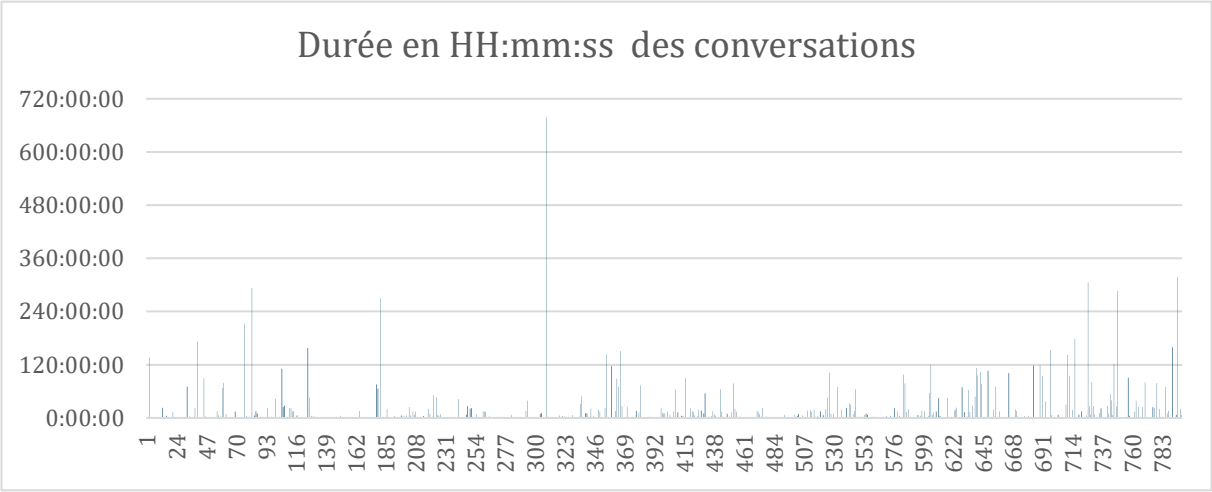


On note des échanges assez courts au cours du début du projet, en temps incrémental qui augmentent davantage pendant la phase de production (sans doute le support, les modifications et corrections) pour culminer à une conversation de 60 messages.

Le nombre moyen d'échange/conversation est de : 3.86



Pour recaler avec le phasage du projet on peut regarder l'évolution de la taille des conversations en temps absolu avec le même plateau que le suivi des messages.

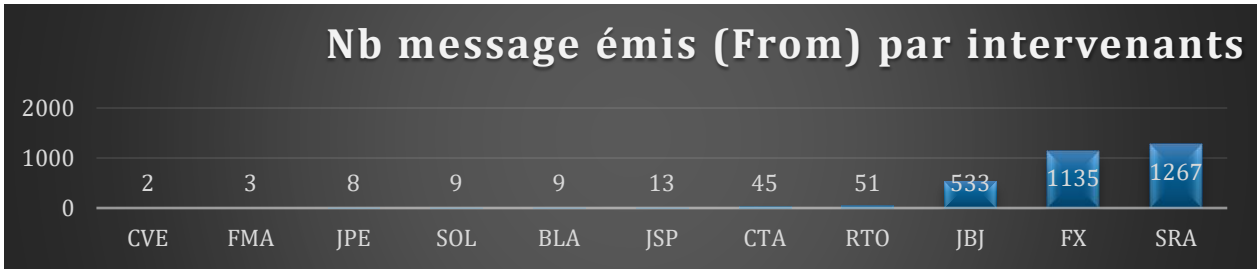


Le calcul de la durée relative des conversations permet de noter que les conversations s'étalent rarement sur plus d'une demi-journée. La durée moyenne d'une conversation est de : 13 :22 :59

6.5.4 La répartition de messages par intervenant

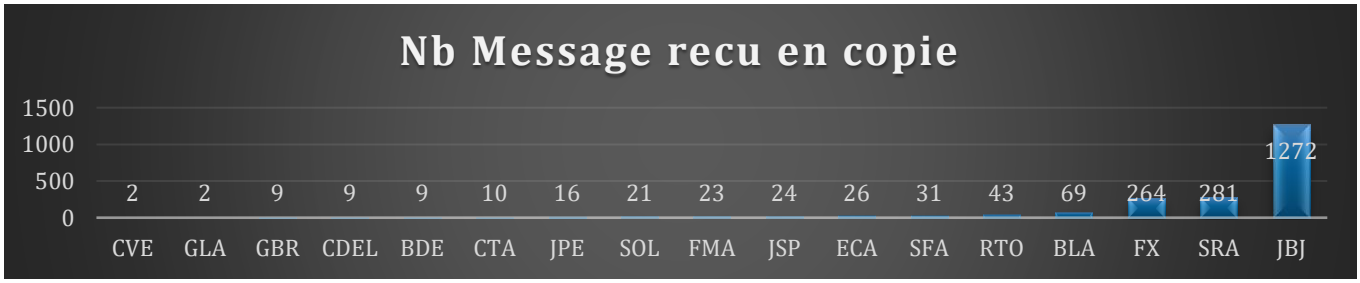
Les intervenants sont des entités reconstruites a posteriori pour mettre en adéquation une seule personne avec plusieurs adresses email (table de correspondance). En effet plusieurs personnes ont changé d'adresse email (ou en ont utilisé plusieurs) pendant le projet.

Nombre de message sortant par intervenant (FROM)



Nombre de message reçus directement (TO) par intervenant

Nombre de message reçus indirectement (CC) par intervenant



Sur ces graphiques, nous avons filtré pour conserver les intervenants ayant au moins 2 messages.

Les graphiques FROM, les TO, les CC permettent de dégager certaines caractéristiques des intervenants :

Le donneur d'ordre (Maitre d'ouvrage opérationnel ou MOA) du projet est celui qui envoie le plus (max des FROM) : SRA

L'exécutant (Maitre d'œuvre ou MO) est celui qui en reçoit directement le plus (max des TO) : FX

Enfin le maitre d'ouvrage officiel (ou institutionnel) est celui qui est le plus en copie (l'observateur invisible) de tout (max des CC) : JBJ

On notera que les grands Responsables selon l'organigramme de la société reçoivent assez peu de mail direct (peut-être marque du respect de la hiérarchie et la délégation des activités).

6.5.5 Graphes sociaux

Nous avons vu dans le chapitre 5, que l'on peut représenter les échanges de mails comme un réseau social. Un réseau social est composé d'acteur (en général des personnes) connectés par des relations (ici l'envoi et la réception de messages) (Scott, 2012).

Les tailles et couleur des nœuds (vertex) sont proportionnelles à leur degré (entrant et sortant), les tailles des arêtes (edges) également. On a appliqué des traitements (lissage des valeurs afin que les nœuds ayant un degré faible reste visibles) pour permettre une visualisation correcte. Le graphe réel est orienté.

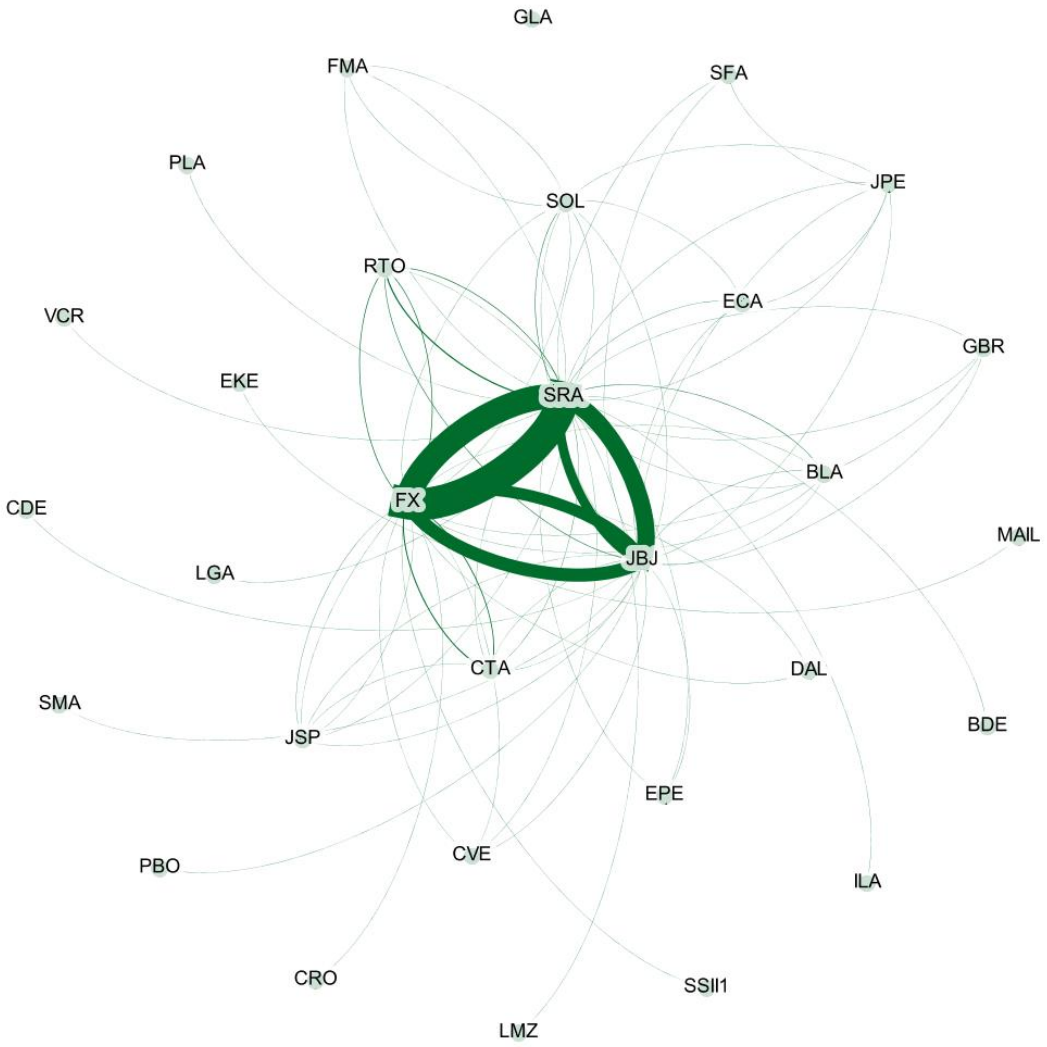


Figure 6.5 : Graphe global des échanges From-To sur l'ensemble des intervenants et la totalité du projet

Le Graphe global des échanges From-To sur l'ensemble des intervenants et la totalité du projet (Figure 6.5) montre la place des 3 intervenants principaux, et surtout les liens importants entre le MO et le MOA.

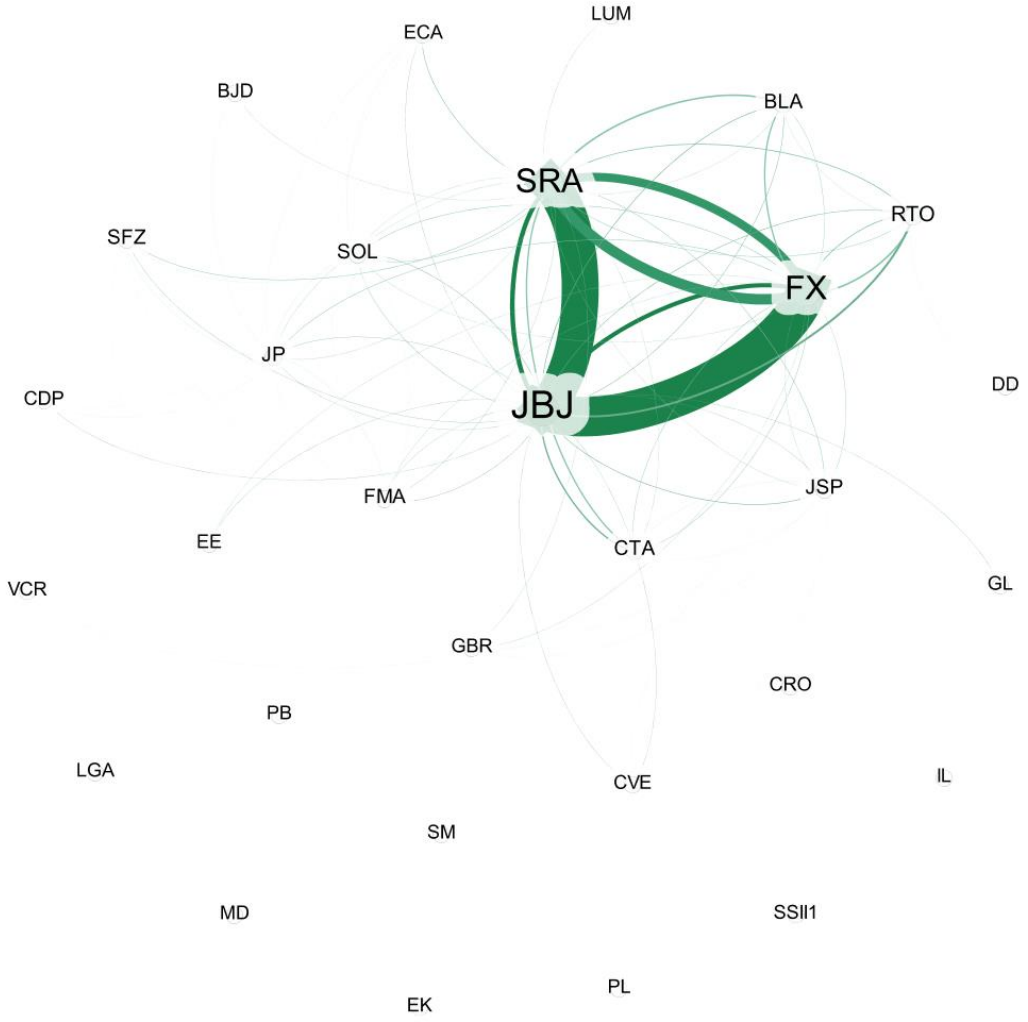
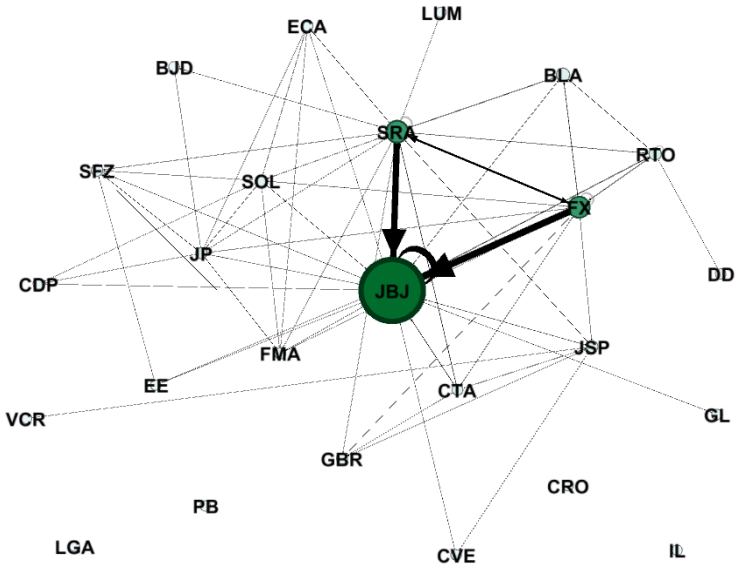


Figure 6.6 : Graphe global des échanges From-Cc sur l'ensemble des intervenants et la totalité du projet

Le Graphe global des échanges From-Cc sur l'ensemble des intervenants et la totalité du projet (Figure 6.6) permet de retrouver bien des flux de « reporting » via les CC (ou comme on peut l'évoquer parfois des « mails parapluies »)

A titre informatif afin de mieux visualiser la direction des arêtes, voici une copie d'écran du graphe « réel » (avant le lissage et la suppression des flèches directionnelle) dans Gephi[19].



On pourrait appliquer des résultats issus de l'analyse de réseaux sociaux (Centralité, intermédiarité, clustering, etc...) pour dégager des tendances et des communautés. Mais le nombre de message (quelle que soit leur nature from, to, cc) est une variable trop simple pour la construction du graphe. Il serait possible mais cela n'entrera pas dans le cadre de cette étude, de procéder à une analyse amont plus fine et plus dirigée pour sélectionner les mots-clefs des sujets et s'en servir dans la construction des arêtes. Enfin, ce mode de représentation laisse de côté une composante essentielle de la dynamique des échanges : la temporalité. Or cet aspect est essentiel dans la résolution de problème, nous pouvons le visualiser en regardant les échanges sur un

[19] https://gephi.org/

thread. Nous faisons figurer ci-dessous une conversation par email entre certains intervenants avec des flèches indiquant un mail TO ou CC et les pièces jointes (PJ). Il est certain que notre approche, basée sur la pragmatique de la communication, doit prendre en compte l'échange dans sa globalité et sa dynamique et non comme un total statique de messages.

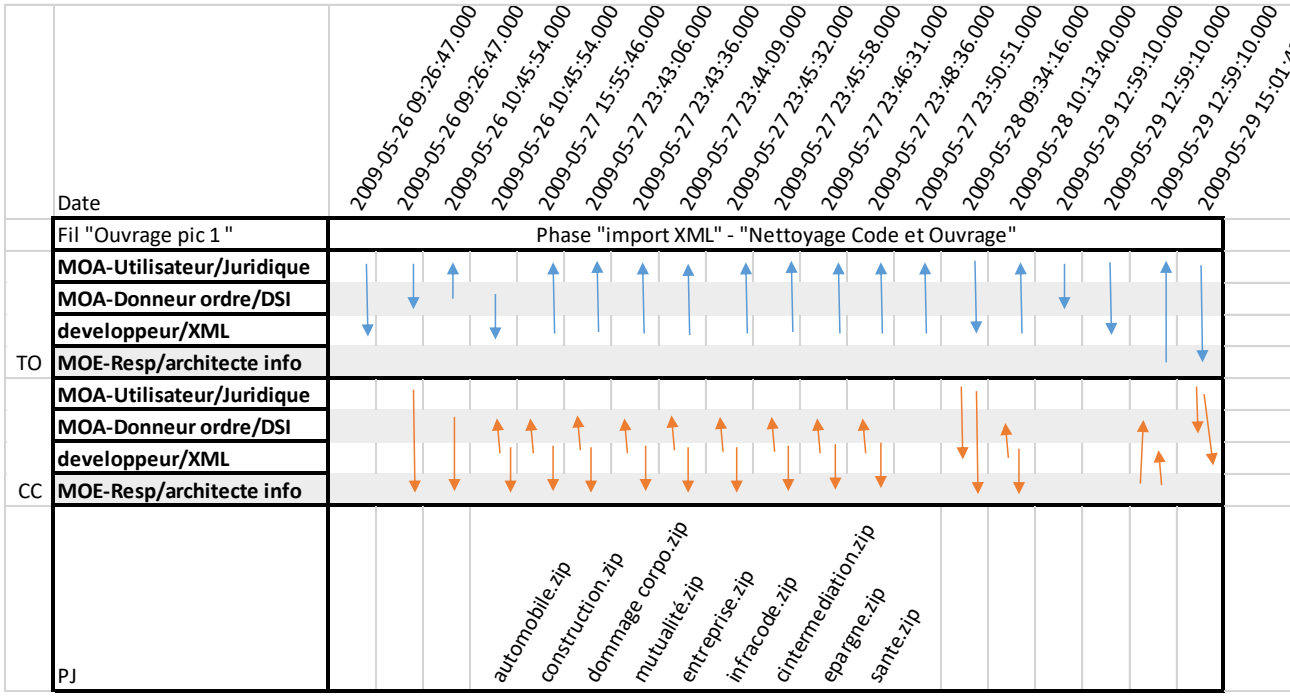


Figure 6.7 : "Conversation" par email

6.6 Annotation

Afin de pouvoir évaluer notre système KTR et mettre en œuvre le module Requête, il a fallu faire une annotation et une labélisation (étiquetage) manuelles du corpus. L'objectif était double : déterminer dans quelle mesure un message contient des traces de connaissance (nous verrons plus loin les restrictions et problématiques que cela soulève) et si le message contient des AL de type requête directe ou indirecte. Ce travail s'intègre dans le cadre de l'analyse de l'activité humaine a posteriori. L'analyse de l'activité passe par un processus d'observation et d'interprétation. Or les messages sur lesquels travaillent nos experts ne sont que des traces de cette activité. Pour comprendre l'activité sous-jacente à ces traces, les experts ont à interpréter ces traces à l'aide de leurs connaissances.

Afin de montrer la pertinence de notre approche, nous avons procéder par une comparaison entre une annotation manuelle des messages et une annotation semi-automatique suivant KTR.

6.6.1 Annotation manuelle

Nous avons dû étiqueter le corpus pour la partie en lien avec l'apprentissage supervisé (SVM pour la requête) et pour l'évaluation finale des performances.

Il n'existe pas de motif (pattern) précis permettant de définir les connaissances ou leur trace utiles pour un projet. Or pour de nombreuses tâches du monde réel, l'annotation manuelle par un expert est le moyen principal d'obtenir les étiquettes. Nous avons gardé une approche pragmatique et orientée monde de l'entreprise. En premier lieu, nous avons demandé à un linguiste d'étiqueter manuellement les messages du corpus pour indiquer la présence de requête directe / indirecte et de noter les phrases. (Un sous-ensemble équilibré a été utilisé pour l'apprentissage du SVM)

Ensuite, il a fallu étiqueter les messages où les réponses à ces demandes avaient été explorées pendant la résolution de problèmes. Dans ce cas, il est impossible d'obtenir l'étiquette réelle (aussi connu comme la vérité terrain (ground truth)) et elle a été estimée à partir de l'opinion subjective d'un petit nombre d'experts.

Pour ce faire, un expert technique sur les domaines du projet a étiqueté les messages qui correspondent à des traces de connaissances collaboratives, et un manager technique qui a travaillé sur ce projet précis il y a 6 ans a bien voulu faire de même.

Bien sûr cette approche est subjective à l'annotateur, mais ces réponses sont très liées à un projet et nécessitent différents domaines d'expertise. L'étiquetage manuel est un processus coûteux et de longue haleine. Donc tous les messages sélectionnés par les deux experts ont été considérés comme de bons candidats pour contenir des traces de solutions aux problèmes.

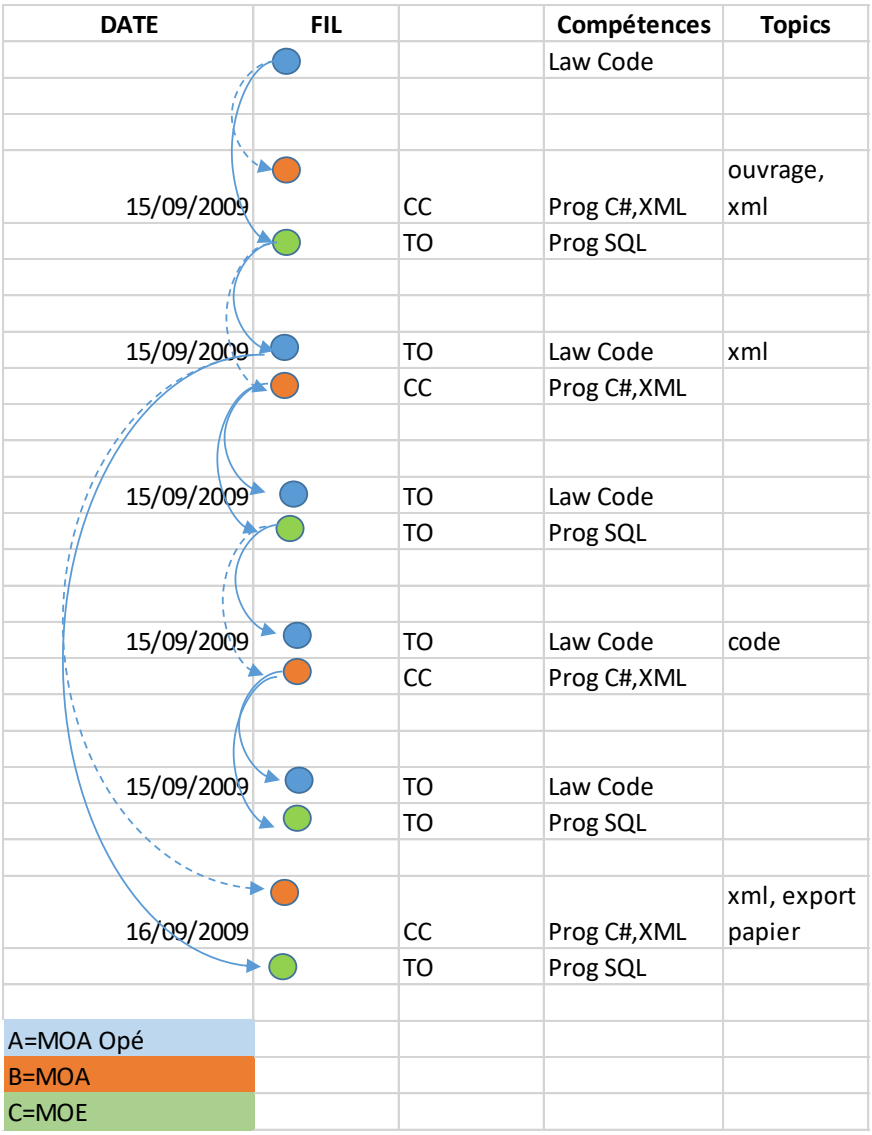
Cette partie a été particulièrement fastidieuse et en raison de contraintes de temps seulement 80% du corpus (à partir du début des e-mails en ordre chronologique) ont été annoté par ces deux experts qui ont marqués 624 messages comme étant porteurs de traces de connaissances.

6.7 Expérimentations

Nos expériences sur le corpus ont été effectuées en plusieurs étapes. Tout d'abord, nous avons bien vérifié sur certains fils de conversation que notre approche globale était correcte. C'est à dire qu'en partant sur des fils de conversation (au moins 3 messages) en rapport avec les sujets du projet, et comportant une requête, nous pouvions trouver des traces de connaissance utiles à capitaliser. Ensuite nous avons mis en place des algorithmes support sur le corpus.

Chronologiquement, une implémentation complétement orientée apprentissage supervisé a été d'abord évaluée, mais elle s'est révélé inadaptée à notre approche et peu convaincante. Ensuite la méthode KTR décrite au chapitre 5 a été mise en œuvre.

6.8 Les conversations



DATE	FIL		Compétences	Topics
			Law Code	
				ouvrage,
15/09/2009		CC	Prog C#,XML	xml
		TO	Prog SQL	
15/09/2009		TO	Law Code	xml
		CC	Prog C#,XML	
15/09/2009		TO	Law Code	
		TO	Prog SQL	
15/09/2009		TO	Law Code	code
		CC	Prog C#,XML	
15/09/2009		TO	Law Code	
		TO	Prog SQL	
				xml, export
16/09/2009		CC	Prog C#,XML	papier
		TO	Prog SQL	
A=MOA Opé				
B=MOA				
C=MOE				

Figure 6.8 : Fil de conversation

Nous visualisons les interactions et les tours de paroles en mettant en évidence les destinataires direct (TO) et ceux en copies (CC). Un exemple d'un tel fil peut être vu dans le tableau figure 6.8 ci-dessus. Cela nous donne un contexte riche pour effectuer une analyse pragmatique. Pour analyser le texte du message et trouver des éléments de résolution de problèmes, nous avons utilisé notre grille personnalisée pragmatique pour identifier les actes de langage de requête. Puis, en suivant le fil des messages par ordre chronologique, nous recherchons des réponses successives à cette requête dont l'émetteur possède les compétences appropriées.

6.9 Première analyse de l'expérimentation

Nous avons testé manuellement notre démarche expérimentale sur quelques fils de message avant de passer à une implémentation des algorithmes. Nous présentons succinctement cette étape afin de visualiser notre approche sur des conversations réelles.

Notre démarche sur une conversation respecte 3 étapes :

- vérifier que les messages ont trait au projet (nous définirons précisément les topics plus bas)
- détecter des requêtes
- regarder si des traces de connaissances sont présentes à la suite des requêtes et observer les compétences des intervenants.

Exemple sur un fil (Figure 6.9) dont le sujet était « textes attachés_the end » :

Date	From	To	Messages	Topics	Requête	Compétence
27/07/2009	SRA cc:	FX JBJ	Les textes attachés du code de la mutualité sont rentrés !!!!!!Hip hip houra ! Pas de problèmes. Petit bémol : pour régler définitivement le dossier texte attaché, vous n'auriez pas vu les textes attachés qui sont nés pendant mes congés (les 26 du code de la Route de GBR) ?...Sinon, on va les traiter manuellement... SRA.	code	0,8	

27/07/2009	JBJ	FX SRA	Je ne peux pas aider, FX, une idée?		0,6
27/07/2009	FX	SRA cc: JBJ	Nous avions les 26 textes de RTO (Word) mais seulement pour le code de la Mutualité (...). Tu trouveras en pièces jointes les textes attachés supplémentaires à rajouter pour le code de la route (sauf le doc intitulé « gratas… »). Il faut les transformer en XML. Soyez attentifs aux balises et au respect du schema xsd ou les feuilles xslt ne fonctionneront pas (...) ps : les annexes seront livrées cette nuit	xml, code	0,75
28/07/2009	SRA	FX cc: JBJ	(..)Les textes attachés de GBR : on s'en occupe à la main dans le Code Automobile. Vous pouvez donc considérer le dossier texte attaché CLOS chez vous. SRA.	code	0,6

Figure 6.9 : exemple sur un fil

Dans cet exemple simple, on peut visualiser les principales étapes de notre démarche.

Dans la colonne « Topics », on retrouvera les sujets en rapport avec le projet. Nous vérifions avec ce paramètre là que les messages sont pertinents.

Dans la colonne « Requête », on trouve une probabilité de requêtes. On voit que des requêtes sont détectées sur les messages 1 et 2.

Dans la colonne « Compétences », on regarde à la suite des requêtes si les utilisateurs qui répondent ont des compétences par rapport au sujet abordé dans le message.

Ici SRA (le MOA juriste) fait une requête directe en tout début de conversation sur le topic Code, JBJ (MOA) va y répondre mais n'amènera aucune informations (pas de topics), FX y répondra et ses compétences sur les sujets abordés dans sa réponse feront que sa réponse sera pertinente du point de vue de notre étude.

En utilisant les données issues de l'étiquetage des experts, nous avons estimé la validité notre hypothèse sur la place de la requête :

- Le nombre de messages contenant des requêtes dans le corpus étiqueté était de 680, dans les fils de conversation de ces messages-là, nous retrouvons 543 messages labellisés comme « porteur de traces de connaissances » par les experts soit près de 87%.

Les valeurs numériques seront explicitées dans l'étape KTR. Ces premiers tests nous ont confortés dans la validité de nos hypothèses de travail et nous sommes passés à l'implémentation des algorithmes.

6.10 Application KTR

Comme indiqué dans l'algorithme en fin de Chapitre 5, la mise en œuvre de la méthode KTR nécessite de préparer plusieurs ensemble de données, structures et matrices associés au corpus. A savoir les Topics (les thèmes du projet), le paramétrage et l'apprentissage du détecteur de Requête ainsi que les Compétences et les rôles des acteurs.

6.10.1 Topics

A parti du phasage réel du projet présenté plus haut et d'un expert technique sur les sujets, nous avons défini notre dictionnaire de Topics. C'est une partie qui est difficilement automatisable, car elle nécessite de prendre en compte à la fois des données projets, mais aussi des données sur les disciplines techniques et également sur les mesure de similarité en IR. En effet, les topics doivent couvrir l'ensemble des phases, et des technologies associées, mais ne doivent pas se recouvrir entre eux (cela fausserait les mesures, idéalement les vecteurs topics devraient être orthogonaux dans l'espace VRM). Ils ne doivent pas non plus être trop « génériques » car sinon tous les messages seraient classés comme ayant trait au projet. C'est donc une heuristique à mettre en place.

Au final, nous avons établi 10 topics avec des mots clefs associés et le dictionnaire était le suivant :

id	label	keywords
1	XML	structuration, balise, arbres, <juri>, <link>, xsd, dtd, schéma, xslt, xml, xmlspy, formatage, xincs
2	BDD	Base, donnée, table, champs, stockage, bdd, identifiant, procédure stockée, select, requête, extraction,sql
3	Interface workflow	UI, Workflow, Interface utilisateur, login, gestion user, asp.net, asp, C#, onglets, javascript, js, arborescence, bouton
4	Code	Code assurance, Legifrance, code automobile, code route, code mutu, mutualité, chapitre, article, bibliographie, commentaire, texte attachés, jurisprudence
5	Ouvrage	nouvelle collection, construction, ouvrage, brochés,

6	Export papier	Indesign, maquette, Christine, mise en page, mapping, balise indd, indd, gabarit, feuille style, réalisation, editing, marges, pagination
7	Export site	export web, balise web, livraison web, dtd web, html, css, xslt web, cross media
8	Export Auteur	word, auteur, xslt auteur, wordml, office
9	Services	update legi, mise à jour Legifrance, FTP, maj legi, synchro
10	Word	macro word, complément, web service, lucene, word 2007, enrichissement, wordlink, add in, lien

Figure 6.10 : Dictionnaire de topics

Avec ce dictionnaire, nous avons calculé la matrice T Avec (Tij) représentant le poids du topics j dans le Message i comme indiqué dans le chapitre 5. Cette matrice nous permet de calculer le score Topic de chaque message.

La figure 6.11 présente un extrait de visualisations sur quelques topics.

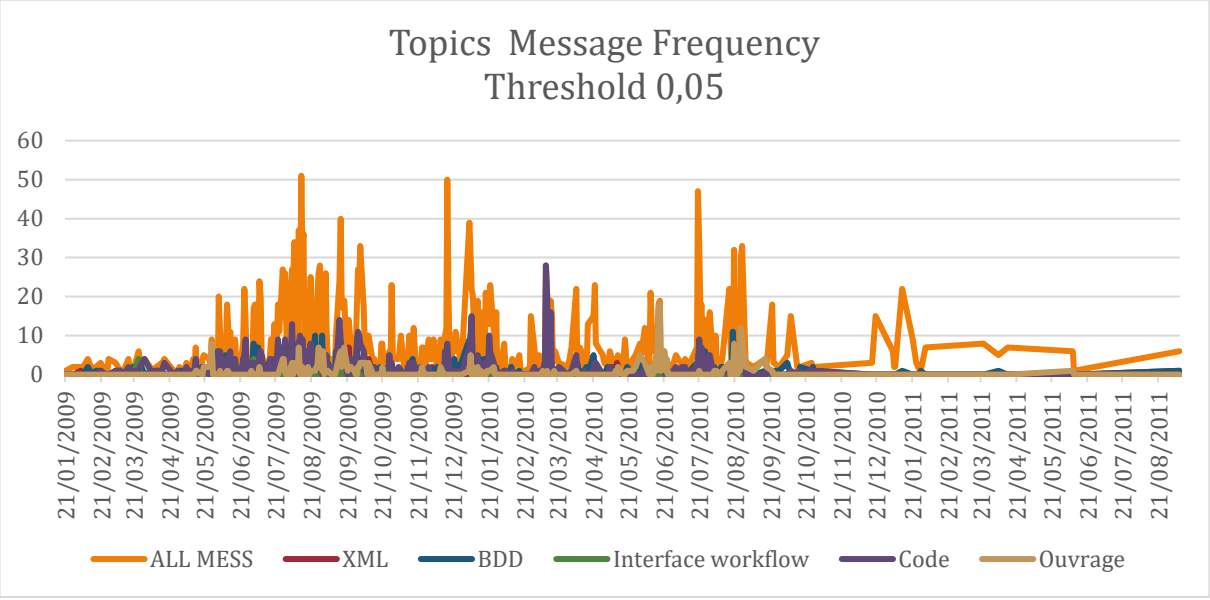


Figure 6.11 : extrait de résultats de l'identification de topics dans les messages

6.10.2 Requête

Afin de calculer le score Requête de chaque message, nous avons employé un Classificateur SVM (Support Vector Machine depuis le Cloud Azure machine ML avec les paramètres par défaut) qui a été entrainé sur 300 phrases (150 comportant des requêtes, 150 sans requêtes).

Pour mémoire le vecteur caractéristique comportait 5 paramètres qui sont normalisés en entrée :

- Présence de signes verbaux (par exemple, pourrait, peut, devrait, pourrait, avez-vous, etc.)
- Présence de mots-clés spécifiques
- Présence marque d'interrogation
- Présence d'une requête détectée dans le même thread (messages précédents)
- Le score d'influence des émetteurs / récepteurs

Les signes verbaux ont été établis avec un linguiste et consistent en des n grammes sur mesure dont voici un extrait :

Verbal_sign ={ peut on, peux tu, il faudrait, aurais tu, nous devrions, il faut, nous souhaitons, pourrais tu, pourriez vous, j'ai besoin, tu pourrais, il serait mieux, on pourrait, voudrais tu, faut il, dois je, il faudra, il manque, est ce possible, est on, puis je, (..)}

De manière analogue pour les mots clefs spécifiques :

Specific_words ={ merci de, question, problème, comment, quand, qui, quoi, combien, bug, erreur, bizarre, dommage, ennuyeux, impossible, soucis, cassé, help, panne, reponse,(..)}

Le score d'influence entre émetteur et récepteurs (pour les requêtes indirectes) est calculé à partir d'une matrice de relation R représentant les liens hiérarchiques et les liens donneur ordre / fournisseurs.

Voici un extrait de la matrice R du projet :

	SRA	JBJ	FX	RTO	BLA	CTA	CVE
SRA	1	0	0,8	0,4	0	1	1
JBJ	0	1	1	0,5	1	0	0
FX	0	0	1	1	0	0	0
RTO	0	0	0	1	0	0	0
BLA	0	0	0,5	0,25	1	0	0
CTA	0	0	0	0	0	1	0
CVE	0	0	0	0	0	0	1

Figure 6.12 : Extrait matrice R

Par exemple JBJ (donneur ordre) a une influence sur BLA (c'est son manager direct) sur FX (c'est son fournisseur direct) et partiellement sur RTO (qui est un sous-traitant du fournisseur)

Il faut noter que ce classificateur n'est pas spécifique au projet et pourra être réutilisé et amélioré, soit par les paramètres du vecteur, soit en augmentant son ensemble d'apprentissage. Le taux d'erreur sur notre ensemble de test est de 0.28. Il serait bien sur possible d'améliorer les résultats ou d'utiliser un autre module (Naïve Bayes ou probabiliste pour la détection des requêtes ou encore prendre le classificateur de Carvalho (2006)) mais ce n'est pas l'objet de notre étude, c'est la combinaison de l'ensemble que nous jugeons significative.

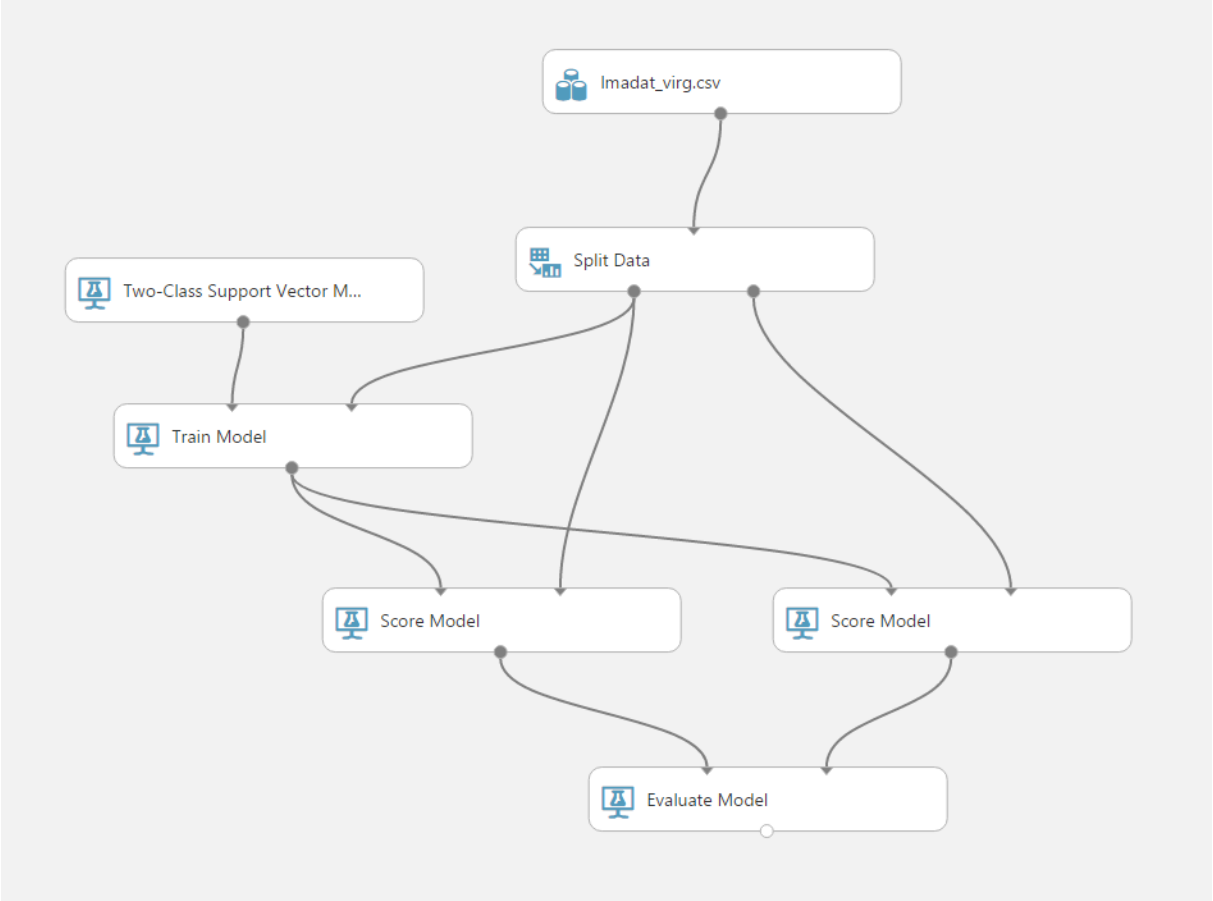


Figure 6.13 : modèle workflow SVM azure ML

Le workflow de modèle SVM dans AZURE ML Studio[20]. On note les données en entrée (lma_dat), la séparation des données d'apprentissage et de test (split data), le SVM (Two-Class SVM), enfin l'apprentissage (train model) et puis les tests.

6.10.3 L'exploitation des Compétences

Pour la dernière partie du score KTR qui vise à établir la capacité des émetteurs à apporter des connaissances à la résolution potentielle de problème en cours dans le fil de conversation.

[20] https://azure.microsoft.com/en-us/services/machine-learning

Pour ce faire nous établissons 2 matrices l'une concernant les compétences des utilisateurs UC, l'autres les compétences utiles selon les topics CT.

La construction de ces matrices se fait simultanément, en effet les compétences utilisateurs que nous allons estimer sont celles utiles dans ce projet.

Là aussi, nous avons sollicité l'expert technique et l'ancien manager, ainsi que les documents de gestion projet afin de déterminer quelles compétences étaient nécessaires pour quelles phases du projet, et à quel degré. Nous avons déterminé 8 compétences techniques en rapport avec les topics du projet.

Voir ci-dessous la matrice CT en page suivante figure 6.14.

	xml	bdd	Interface workflow	Code	Ouvrage	papier	Export site	Export Auteur	Services
prog XML	1	0	0		0,7	0,5	0,5	0,5	0
progC#	0,25	0	1	0		1	1	0,5	1
prog SQL	0	1	0,5	0	0	0	0	0,125	0,5
Architecture	0,5	0,5	0,5	0	0	0	0	0	0
law code	0	0,13	0,5	1	0,7	0	0	0	0
law writing	0	0	0	1	0,7	0	0	0	0
indesign	0	0	0	0	0	1	0	0	0
HTML	0	0	0,5	0	0	0	1	0	0

Figure 6.14 : Extrait matrice CT

Ensuite nous avons consulté la Direction des ressources humaines du groupe et les documents projet avec les CV des sous-traitants.

Voir ci-dessous un extrait (pour les principaux intervenants) de la matrice ^{t}CU

	prog XML	Prog C#	prog SQL	Architecture	law code	law writing	indesign	HTML
SRA	0	0	0	0	1	1	0,125	0
JBJ	0	0	0,125	0	0	0	0	0
FX	1	1	1	1	0	0	0,5	1
RTO	1	0	0	0	0	0	0	0,125
BLA	0	0	0	0	0	0	0,125	0,125
CTA	0	0	0	0	1	1	0	0
CVE	0	0	0	0	0,5	0	1	0

Figure 6.15 : Extrait matrice ᵗCU

Cette matrice est intéressante car elle dépasse les rôles projet, en effet par exemple sur JBJ, le manager du projet coté DSI, son rôle était plutôt d'orchestrer le bon déroulement du projet, de planifier, prendre des décisions, donc plutôt des « softs compétencies », cependant son CV mentionne qu'il a travaillé sur les bases de donnée pendant de nombreuses années. Aussi la matrice en tient compte car les réponses qu'il pourra faire par email sur ce sujet pourront être pertinentes.

On peut voir par exemple que l'utilisateur SRA possède de bonnes compétences dans le domaine juridique et une connaissance d'Indesign, et que ces compétences sont importantes pour les tâches des Topics Code et papier. Les valeurs des compétences ici sont arbitraires et servent seulement à calculer un score de correspondance avec les Topics.

Avec ces matrices comme indiqué au chapitre 5, on obtient la matrice UT ($UT = {}^{t}CU.CT$) avec (UT_{ij}) représentant une estimation approximative des compétences de l'utilisateur i en ce qui concerne le Topic j.

Ci-dessous en figure 6.16 un extrait de la matrice UT (avant normalisation)

	xml	bdd	Interface workflow	Code	Ouvrage	papier	Export site	Export Auteur	Services	Word
SRA	0	0,13	0,5	2	1,4	0,13	0	0	0	0
JBJ	0	0,13	0,0625	0	0	0	0	0,02	0,06	0,0625
FX	1,75	1,5	2,5	0	0,7	2	2,5	1,13	1,5	2,1
RTO	1	0	0,0625	0	0,7	0,5	0,63	0,5	0	0,8
BLA	0	0	0,0625	0	0	0,13	0,13	0	0	0
CTA	0	0,13	0,5	2	1,4	0	0	0	0	0
CVE	0	0,06	0,25	0,5	0,35	1	0	0	0	0

Figure 6.16 : Extrait Matrice UT

Avec le corpus étiqueté par les experts, nous avons examiné les messages classés comme « traces de connaissances » et nous avons regardé si les émetteurs avaient des scores pertinents dans les topics majeurs des messages. Pour ce faire, nous avons sélectionné les 2 premiers topics ayant le plus haut score dans ces messages, et nous avons regardé si dans la matrice UT ci-dessus, les émetteurs avaient des scores significatifs (nous avons définis un seuil à 0.5). Dans 67 % des cas, les connaissances des messages émetteurs sont liées à leurs compétences sur les sujets. Cependant, nous verrons que le score final tient compte d'un contexte plus global. En outre nous avons remarqué en examinant des fils particuliers que cela dépendait bien sûr de la manière dont les compétences nécessaires étaient attribuées aux topics mais aussi que les utilisateurs pouvaient, en collaborant avec l'équipe, dépasser leurs compétences « initiales » (selon les rôles dans le projet). Pour être précis sur un exemple,

l'utilisateur SRA dans un fil d'Avril 2009 va proposer une modélisation XML différente qui sera adoptée au final et pourtant c'est un juriste.

6.10.4 Indexation et Recherche

Une fois la mise en place de ces données préparatoires, on peut calculer les sous scores et le KT_Score sur l'ensemble des messages. La phase d'indexation est terminée (première étape de l'algorithme KTR décrit fin de chapitre 5). Dans une deuxième phase, la recherche suivant des requêtes utilisateur Q est alors testée en combinant le KT_score avec la similarité message/requête Q.

Le classement final d'un message M_i étant obtenu par combinaison linéaire :

$$r(M_i) = \mu Score\_KT(Mi) + (1-\mu)\text{sim}(M_i, Q)$$

6.10.5 Réglages de paramètres

Pour le réglage fin des paramètres, c'est encore une heuristique basée sur le savoir des experts qui intervient. Par exemple certains mots-clefs du dictionnaire des topics sont « boostés », c'est-à-dire que l'on augmente leur importance dans le score final (par exemple « xml » ou « balises » sur le topic XML). De même pour les sujets des emails qui ont plus grande pondération. Certains ajustements (scaling) ont été faits pour ramener les valeurs des sous scores sur des échelles comparables.

Le réglage du paramètre de combinaison μ a été fixé à 0.40 après nos tests sur les requêtes (voir partie résultats plus bas)

6.10.6 Résultats

Une visualisation de résultats est d'abord avancée avant de discuter des performances de l'approche KTR.

6.10.6.1 Visualisation

Afin de mieux visualiser les scores de la méthode KTR, nous avons fait figurer les résultats sur un fil exemple en page suivante.

FROM	TO	CC	Message	Topics	Topics Part	Request part	Compétencies part	KT_Score
SRA	FX	JBJ	J'ai un ENORME doute tout à coup...Je suis en train de faire le mail à LAR et j'ai l'impression que c'est nous qui n'avons pas compris !(..) - Et pas un contrôleur de statut des articles => ce n'est pas lui qui nous indique si on doit afficher ou pas un article. c'est la balise <ETAT> le .dat, c'est pour préserver iron dans 10 ans (pas trop de fichiers)... CQFD. Etes vous ok avec ça ?	XML,Code	0,07	1		1,07
FX	SRA	JBJ	il vaut mieux comparer les codes direct, les maj sont parcellaires, et notre probleme et que l'on a appliqué une maj complete (géante) (..) sur un code qui devait etre incorrect en partie. (..) mais en ayant une liste un fichier texte des id legi (arti+scta) a "abroger"/supprimer logiquement, deja on pourrait se débarasser des branches xml (scta) et feuilles surnuméraire. 2 missions :- la réparation du code actuel-la verification du process d'abrogation(plus de dérive future)	XML,Code	0,05	0,5	0,1025	0,6525
JBJ	SRA,FX		Moi j'avais compris ca....		0	0	0	0
SRA	JBJ,FX		T'es le meilleur		0	0	0	0
SRA	FX	JBJ,JSP	on, mission 1 : terminée de notre côté => ci-joint l'audit sur le code des assurances.(..)Bonne nouvelle : on repère des erreurs récursives. Mauvaise nouvelle : je ne vois pas où est le problème ?(à part sur les abrogations : (..)Le problème est que je ne vois pas pourquoi la livraison du 03/03 aurait posé des soucis, vu qu'elle n'a rien changé (précisément...) Et que la maj du 12/03 est tout ce qu'il y a de plus banal....	Auteur	0,08	0,5	0	0,58
FX	SRA	JBJ,JSP	(..) 1)J'ai recodé les abrogations en tenant compte de la nouvelle approche (cad abrogation pas dans liste-suppression_legi.dat) =>je regarde les scta et article etat "ABROGE" dans l'arbre en vigueur (je les abroge logiquement s'ils sont encore actifs) Après Tests => j'ai rajouté une autre routine, (...) et alors je comprends pourquoi l'abrogation de l'article R411-1 n'a pas eu lieu.. si tous les noeuds pere sont abrogés, le bot d'exploration des mises a jour ne va pas explorer les branches mortes(abrogées) (puisque qu'on ne les voit pas sur l'arbre', donc ne recopie jamais le nouveau fichier LEGISCTA000006176933. mais dans le Worfkflow ca devrait marcher car comme la branche père est abrogée, elle n'apparait pas dans l'arbre.. les branches C1,C2,C3 de Titre premier de L4 ne devrait donc pas apparaitre si on en croit LEGISCTA000006143456 du 12/03 (toute en etat "ABROGE")	XML,Code,Interface	0,12	0,5	0,2275	0,8475
SRA	FX	JBJ,JSP	1° ok. 2° ok. C'est ça . les branches C1,C2,C3 de Titre premier de L4 ne devrait donc pas apparaitre si on en croit LEGISCTA000006143456 du 12/03 (toute en etat "ABROGE") Ça te fait du vrai curatif. 3° ok. ça doit être un effet de bord du au chargement du stock en maj. 4° ça devrait marcher ... la juri est plus					

6.10.6.2 Performances

Pour mesurer les performances en recherche (IR), nous avons décidé de comparer KTR avec la Cosine Similarité(CS) basique (basée sur les TF IDF) (recherche de texte Lucene).

Ceci a été réalisé en réglant le paramètre μ à zéro. Nous avons forgé 20 requêtes sur la base de scénari plausibles de l'entreprise (par exemple, « code assurance format de fichier xml ») et comparé les résultats des deux méthodes en comptant le nombre de résultats marqués pertinents par les experts technique et projet dans les 50 premières réponses (ou le maximum du nombre réponses si celui ci était inférieur).

Nous avons constaté une amélioration de 8% sur les mots-clés seuls sur nos requêtes de tests.

	Nb Message CS	Nb Messages KTR	%
10 requêtes	38	42	11%
20 requêtes	83	90	8%

Nous avons ensuite regardé l'amélioration sur les 20 premiers résultats du classement afin de déterminer si le score KTR faisait « monter » les messages pertinents dans le classement.

20 premiers	Nb Message CS	Nb Messages KTR	%
10 requêtes	28	33	18%
20 requêtes	57	66	16%

6.10.6.3 Commentaires des résultats

Ces résultats appellent des remarques et commentaires :

- Tout d'abord, il faut se rendre compte que c'est une amélioration à relativiser, cela fait 1 résultat pertinent de plus par requête en moyenne. En revanche ce qui notable, c'est que cela améliore le classement global, et c'est intéressant pour une application pratique.
- Bien sûr, ce sont des résultats préliminaires et il nécessiterait beaucoup plus de tests et de mesure (comme les mesure précises F1 et faux positif / négatif, et). Les experts ne sont pas facilement disponibles étant donné que notre corpus est issu du monde réel, mais nous aimerions travailler avec eux sur les résultats en matière d'étiquetage basés sur les requêtes.

En effet, l'évaluation des performances est difficile dans ce cas-là, car idéalement, il faudrait que sur chacune des requêtes type, l'on puisse passer en revue l'ensemble des messages du corpus pour noter les plus pertinents. Cela reviendrait à demander à l'expert de revoir 20 fois les 3080 messages pour décider s'ils sont pertinents ou pas. C'est une tâche très importante que nous n'avons pas pu mener à bien dans cette étude.

Les experts ont annoté une fois le corpus en regardant tous les messages et en décidant s'ils étaient porteurs de trace de solution à un problème et de connaissance dans l'absolu (car ils connaissaient le projet ou les domaines techniques) et non relativement à une requête précise. Néanmoins sur des fils exemples, nous avons pu constater que les messages ayant un KT_score important étaient en général intéressants d'un point de vue de la traçabilité. L'évaluation des performances dépend aussi des requêtes, plus une requête est précise, moins elle ramènera de résultats et donc les gains seront minimes entre CS et KTR.

Notre partie requête pourraient à être améliorée sur deux aspects :

- Les requêtes indirectes
- La notion de sujet (Topics) de « voisinage » de requête

Tout d'abord les requêtes indirectes sont très difficiles à détecter. Pour prendre un exemple sur un fil :

SRA va émettre un message le 15/09/2015 contenant « Si je pouvais mettre du <juri> avec tout ce qui va avec dans les ouvrages, ça serait pas de refus…(..) », comme il écrit au fournisseur, c'est véritablement une requête pour un potentiel développement qui va donner lieu à un échange d'email sur les modalités puis à des spécifications. Seulement la requête n'est pas détectée (pas de signe d'interrogation, ni de verbe ou de mots clef..) et donc les messages de cette conversation auront seulement des scores bien en dessous de leur véritable « valeur ». Des pistes pour remédier à cela seraient de prolonger l'apprentissage du SVM et d'éventuellement l'enrichir de nouvelles caractéristiques.

Ensuite une limitation inhérente à la forme actuelle de notre algorithme : Nous détectons les requêtes à la granularité de la phrase et les Topics à celle du message. Ce qui peut mener dans certains cas à une forme de « faux positif» sur les requêtes. Nous avons bien une requête détectée et effectivement sur la forme, c'est bien une requête, et elle se trouve dans un message qui contient beaucoup de topics du projet, seulement la requête ne les concerne pas : Un exemple extrême (mais tiré du corpus) : à la suite d'une présentation des prochains développements XML, l'émetteur demandait au récepteur s'il voulait déjeuner à midi. Ces détections font remonter le score KTR de messages qui ne devrait pas être mis en avant car ne faisant pas partir d'une potentielle résolution de problème. Cependant on ne peut se fier uniquement au topics de la phrase contenant la requête, il faudrait donc mettre en place une notion de voisinage (au sens distance de mots) autour d'une requête détectée dans un message pour repérer plus finement les topics de la requête et par là même s'en servir dans la partie solution (chercher les messages et les compétences en rapport avec les topics de voisinage de la requête).

Ces résultats sont prometteurs, mais loin d'être suffisants pour un usage direct dans le monde de l'entreprise.

6.11 Conclusion

Au final, l'expérimentation sur un corpus d'entreprise s'est révélée très enrichissante.

Elle nous a tout d'abord, et ce, déjà à travers les travaux d'experts permis de confirmer que les messages emails contiennent bien des traces de résolution de problème de logique de conception, et donc de connaissances et pouvaient être utiles pour la mémoire de Projet.

En second lieu, la pragmatique et la requête se révèle efficace pour détecter ces zones, et enfin il est vrai que lorsque les compétences de l'utilisateur interviennent en aval dans le fil, les messages sont très susceptibles de comporter des informations techniques pertinentes (il est vrai que notre corpus était assez technique axé sur le droit/juridique et la conception de logiciel).

En revanche, l'automatisation de cette démarche s'est avérée bien délicate. Notamment en ce qui concerne l'annotation et l'étiquetage. Il est vrai que si le système KTR est bien « rodé » (apprentissage SVM et tuning du paramètre μ). Cet étiquetage ne sert alors qu'à mesure les performances. Mais dans notre évaluation, la phase d'annotation était cruciale et elle s'est révélée longue, complexe, et subjective. De même la définition des requêtes de test pourrait faire l'objet d'une étude en soi.

Néanmoins cette implémentation nous a permis d'écarter une piste apprentissage automatique complet, de mieux comprendre les corpus de messages professionnels et de mettre en application les fondements de notre méthodologie. Nous ne tirons les leçons suivantes :

L'approche du contexte « global », c'est-à-dire projet, humain, et textuel est pertinente pour la recherche de traces de connaissances dans des corpus de messages.

La pragmatique de la communication est une approche très intéressante et particulièrement adaptée à l'étude des emails où les intervenants sont bien identifiés (contrairement à un forum ou à une conversation orale enregistrée) avec des caractéristiques précises. Ceci permet d'aller plus loin qu'une simple analyse textuelle et de progresser vers les intentions des interlocuteurs.

L'ingénierie des connaissances est précieuse dans ces aspects-là, car elle nous fournit le cadre indispensable à notre analyse du projet. L'application d'hypothèses et d'une méthodologie avec la réalité d'un corpus d'entreprise est toujours un moment que l'on anticipe un peu. Ce contact avec le réel est pourtant riche d'enseignements car s'il nous montre les écarts entre la réalité et le modèle, il nous indique aussi les différents moyens de les combler et par là même de mieux appréhender (comprendre) cette réalité dans son fonctionnement intime.

Cette première expérimentation sur un corpus en nous montrant les limites des automatisations nous a ouvert également de nouvelles perspectives tant au niveau méthodologique qu'à celui du champ exploratoire sur lesquelles nous reviendrons dans nos conclusions de cette étude dans le chapitre suivant.

Chapitre 7. CONCLUSIONS

Au terme de cette étude, nous commencerons par un bilan de notre recherche au regard des objectifs établis en introduction, puis nous examinerons la validité de nos hypothèses de travail. Enfin nous résumerons nos apports et présenterons les perspectives.

7.1 Bilan de recherche

Piaget définissait l'épistémologie comme « l'étude de la constitution des connaissances valables », ce qui amenait les questions : qu'est-ce que la connaissance, comment est-elle constituée, et comment apprécier sa valeur. Dans ce travail, nous avons essayé dans le cadre particulier des courriels d'entreprise concernant un projet de suivre un questionnement similaire : Ces emails contiennent t- ils des connaissances, des traces de génération de cette connaissance et comment l'exploiter et la valoriser ? Notre apport générique est donc double : les emails de projets contiennent des traces exploitables de connaissance et des méthodes (comme KTR) permettent de mieux les retrouver.

Plusieurs chercheurs en gestion de connaissance défendent le postulat que la connaissance est créée à travers une interaction homme/homme ou homme/outils (Grundstein, 1996, Nonaka, 1995). Nos travaux s'intègrent dans cette mouvance. Nos questions de recherche étaient essentiellement basées sur : Est-ce qu'une interaction

Homme/Homme à travers un outil tel que la messagerie électronique peut témoigner de cette création de la connaissance. Dans nos études, nous avons exploité plusieurs techniques allant des approches s'apparentant au : - Traitement du langage naturel et l'ingénierie des connaissances pour déterminer les concepts d'un domaine, - pragma-linguistique afin de reconnaître les intentions des interactions médiatisées, et- enfin la gestion des connaissances et notamment la mémoire de projets pour reconstruire le contexte de production de cette connaissance. Nous démontrons à travers les lignes de ce rapport l'apport de ces techniques dans la reconnaissance de traces des connaissances dans les interactions médiatisées. Notre approche d'analyse est pragmatique partant d'un problème professionnel au sein d'une entreprise. Nous présentons dans ce qui suit les étapes de notre démonstration.

Dans un corpus de messages professionnels, avec les données de contexte du projet (intervenants, organisation, document, produit résultat), nos objectifs initiaux présentés en introduction étaient les suivants :

Objectif 1

Notre premier objectif était de déterminer s'il était possible de localiser des traces de « connaissances » dans ces emails utiles pour la Mémoire de Projet (Matta, 2001). Cet objectif a été atteint, les emails professionnels contiennent des informations utiles pour la mémoire de projet, informations qui dépassent l'usage usuel de planning ou de reporting des emails.

Objectif 2

Notre second objectif à la suite du premier était de déterminer quel type de traces de connaissance étaient localisables et comment. Nous sommes focalisés sur la trace de résolution de problème en Logique de conception et avons choisi la pragmatique comme voie médiatrice.

Objectif 3

Notre troisième objectif était d'établir si ce procédé était automatisable et d'examiner les possibilités pratiques de mises en œuvre dans le monde de l'entreprise. Cet objectif a été partiellement atteint, nous avons présenté au chapitre 5 une méthodologie KTR permettant de faire de la recherche sur les messages

professionnels en mettant l'accent sur les traces de résolution de problème. Ceci démontre les limites des approches automatiques actuellement défendues comme l'identification de termes et de syntaxes spécifiques dans les e-mails. Les performances de cette méthode montrent des améliorations par rapport à une recherche standard mais demande des perfectionnements voire un travail de préparation spécifique pour une utilisation dans le milieu de l'entreprise.

7.2 Bilan des hypothèses

Afin de statuer sur nos hypothèses de travail, nous devons répondre aux questions suivantes :

- Les emails d'un projet contiennent-ils des traces de connaissances utiles à l'entreprise ? Nous avons explicité dans les chapitres 1 et 2 en quelle mesure la traçabilité de la logique de conception peut apporter des informations cruciales notamment dans le cycle de la vie du produit et son évolution. Cette hypothèse s'est vérifiée sur nos analyses de fils de conversations au chapitre 6.
- La pragmatique de communication et notamment l'AL Requête couplée aux données du projet peut-elle nous aider à localiser les séries de messages contenant des éléments de résolution de problème ? Nous avons vérifié au chapitre 6 sur notre corpus annoté par les experts que cette hypothèse semble correcte.
- La prise en compte des compétences utilisateurs enrichit elle la prise en compte du contexte global ? Là aussi, en examinant les messages classés comme « trace de connaissances » par les experts au chapitre 6 en regard des compétences des émetteurs, un lien existe mais il n'est pas significatif en lui-même, il devient intéressant en l'associant à un score global.

7.3 Bilan de la mise en œuvres des algorithmes supports

Dans ce travail, nous avons proposé le système KTR pour analyser les e-mails professionnels d'un projet afin d'aider les utilisateurs à trouver des traces de résolution

de problèmes. Notre étude vise à la traçabilité dans le contexte de la mémoire du projet. Contrairement aux modèles précédents visant à extraire des éléments significatifs (comme des tâches ou des concepts, etc. ...) à partir des e-mails en utilisant des techniques NLP, nous avons insisté sur l'amélioration de la prise en compte du contexte et l'intégration de la pragmatique et de composantes organisationnelles (actes de langage, la modélisation des compétences techniques, des rôles). Les résultats de cette analyse reflètent la richesse en termes d'aide à l'interprétation des traces reconnues pour une construction d'une connaissance ancrée dans une dynamique d'actions.

Objectifs atteints par KTR

Les résultats numériques montrent que la tâche de détection de traces de connaissances de résolution de problèmes est très délicate et spécifique à une activité donnée. Cependant, cette tâche est nécessaire et est accomplie quotidiennement par des employés dans la vie professionnelle. Le système KTR apporte une légère amélioration dans le classement des messages recherchés dans un corpus. Il prend en compte des éléments de contexte humain et organisationnel qui dépassent la simple recherche full text.

Limites et difficultés de KTR

Nous avons vu dans notre corpus exemple que certaines traces de connaissances de Logique de conception ne provenaient pas forcement de requêtes initiales et les demandes et les réponses sont souvent entrelacées créant un contexte bruyant pour les traitements textuels. De plus les problèmes concernant la granularité (niveau phrase ou message) de notre approche nécessiteraient une étude plus approfondie pour parvenir à une meilleure compréhension des modes d'échanges lors de la création de connaissances collaborative. Enfin lors de la présentation des résultats à l'utilisateur, les parties « pertinentes » sont parfois mélangées dans les messages longs avec des informations non pertinentes et soulèvent également des problèmes de vie privée.

Au cours de l'étiquetage expert, nous avons remarqué les limites de notre approche basique de modélisation des compétences: des solutions pertinentes ont parfois été apportées par des personnes qui n'ont pas d'après leur CV des compétences fortes sur

les sujets et/ou qui possèdent des compétences non-techniques (comme la gestion, la planification). Un autre facteur important est l'aspect temporel: décider si le message fait partie de la solution à un problème nécessite souvent de prendre en compte plusieurs fils de conversation.

Ces réserves sur les capacités actuelles de KTR ne bloquent cependant pas l'avancée de nos travaux, bien au contraire, elles nous ont permis de mettre en avant des points d'analyse du corpus à approfondir et de définir des fonctionnalité nouvelles dans les modules. Nous reviendrons sur ces aspects dans les perspectives.

7.4 Apports de notre travail

Notre travail propose plusieurs apports pour la recherche, en termes de méthodologie et de résultat.

Nous avons proposé dans cette étude :

Une méthodologie prenant en compte un contexte que nous qualifions de plus *global* afin d'analyser un corpus de message pour y repérer des traces de connaissance. En effet, notre contexte prend en compte les données du projet, les données textuelles mais utilise en outre la pragmatique et les rôles et compétences des utilisateurs.

En terme de résultats obtenus, nous avons proposé KTR un système de recherche d'information « IR » qui calcule un score sur des messages à partir des données sur les utilisateurs, le projet et la pragmatique et pourraient inspirer d'autres travaux comme l'analyse des communautés, de réseaux sociaux, etc.

7.5 Perspectives

Les apports précédents pourraient être prolongés de plusieurs manières.

A l'échelon de la méthodologie :

Il serait intéressant d'investiguer sur d'autres types de trace de connaissances (autre que la résolution de problème) en combinant notre méthode avec d'autres actes de langages que la requête, et donc d'autres grilles d'analyse. Par exemple, nous expérimentons notre approche sur des messages extraites d'une communauté de pratique afin d'identifier le phénomène d'apprentissage dans ce type d'interactions.

Au niveau des algorithmes :

Notre système KTR étant modulable, il serait judicieux de le tester avec d'autres systèmes plus perfectionnés de détection d'acte de langage. En ce qui nous concerne, nous explorons des pistes en élargissant le contexte aux messages environnants (inter-fils) et en prenant en compte le contenu de pièce jointe. Comme évoqué en fin de Chapitre 6, nous visons également une correspondance plus fine entre les sujets des requêtes et les réponses possibles. Enfin, en explorant des techniques non supervisées d'IA pour les modules de classification.

Pour l'exploitation des compétences, il faudrait explorer d'autres modélisations, par exemple un côté dynamique des compétences. En effet dans notre modèle, lorsque nous associons des compétences à un Topic, ceci est effectué de manière statique or les compétences nécessaires à un sujet varie selon les étapes, une granularité plus fine des tâches pourrait se révéler plus adaptée.

Enfin d'un point de vue purement technique, on pourrait regarder d'autres mesures de similarité à la place de TF-IDF comme KL- divergence, l'expansion de la requête (Tao, 2006) ou encore les parcours de graphes (Blanco 2007).

Enfin au niveau des champs applicatifs :

Nous envisageons bien entendu de l'appliquer à d'autres corpus projets respectant nos critères mais il serait possible d'évaluer un scénario où le projet n'est encore pas terminé et impliquerait les utilisateurs, voire travailler sur une traçabilité au quotidien de l'activité (Matta et al, 2016).

Une manière différente d'appliquer notre méthodologie serait l'évaluation des compétences en entreprise. Généralement elle est accomplie par des entretiens annuels avec les managers ou du personnel des Ressources Humaines. On pourrait utiliser une partie de notre méthodologie afin d'évaluer quelles compétences, aptitudes ont été mises en œuvre par les intervenants sur un projet en se basant sur les traces dans leur emails. La prise en compte des capacités organisationnelles et « softs competencies » (Rosas, 2009) serait à étudier.

7.6 Conclusion

Notre travail constitue donc une exploration dans un domaine à la frontière de l'ingénierie des connaissances, de la pragmatique de communication et du traitement informatique du langage naturel.

Le principe de notre approche est de prendre en compte tous les aspects des échanges médiatisés au cours d'un projet. Ainsi, nous avons présenté une ligne directrice générale pour détecter les discussions pertinentes (liées à la résolution de problème) dans une grande quantité de messages et nous l'avons appliquée dans le cadre du développement logiciel. Pour illustrer cela, nous avons analysé un corpus de projet professionnel et détecté certaines parties de problèmes et les réponses possibles. Sur l'ensemble, ces résultats confirment l'applicabilité de la technique, mais aussi révèlent certaines limites, surtout dans l'automatisation. Cependant cela nous a conforté dans l'idée que les emails professionnels peuvent être exploités de manière à alimenter le capital connaissance d'une entreprise et ne se cantonnent pas simplement un aspect communication. Il également apparu que la prise en compte du seul contenu textuel du courrier électronique (et non le contexte projet et humain dans son ensemble qui l'entoure) conduirait à des résultats très limités en ce qui concerne la détection de traces de connaissances collaborative. En fin de compte, notre travail contribue à la traçabilité de la mémoire du projet et la structuration des connaissances dans la réalisation quotidienne du travail de projet.

BIBLIOGRAPHIE

A

Aggarwal, Charu C., and ChengXiang Zhai. Mining text data. Springer Science & Business Media, 2012.

Androutsopoulos, Ion, et al. "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach." arXiv preprint cs/0009009(2000).

Antoniol, G., Canfora, G., Casazza, G., De Lucia, A.: Information retrieval models for recovering traceability links between code and documentation. In: Proceedings of 16th IEEE International Conference on SoftwareMaintenance, pp. 40–51. IEEE CS Press, San Jose, CA (2000a)

Asuncion, H. U., & Taylor, R. N. (2012). Automated techniques for capturing custom traceability links across heterogeneous artifacts. In Software and Systems Traceability (pp. 129-146). Springer London.

Atifi H., Gauducheau N., Marcoccia M.2011. The Effectiveness of Professional Emails: Representations and Communicative Practices , in proceedings of 13th Conference of the International Association for Dialogue Analysis, Dialogue and Representation, Montréal.

Atifi, Hassan, and Michel Marcoccia. (2006) "Communication médiatisée par ordinateuret variation culturelle: analyse contrastive de forums de discussion français et marocains." Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires 9 (2006): 59-73.

Atifi, Hassan, Nadia Gauducheau, and Michel Marcoccia. "Les manifestations des émotions dans les forums de discussion." Journées d'étude «Emotions et interactions en ligne», ICAR ENS LSH/Lyon 2 (2005).

Aussenac-Gilles N., Matta N., Making a method of problem solving explicit with MACAO, in International Journal of Human-Computer Studies, Vol. 40, 1994.

Austin J. L. (1962), How to do things with words. Harvard University Press, Boston, MA, 1962.

B

Bacchelli, A., Lanza, M., Robbes, R.: Linking e-mails and source code artifacts. In: Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering, vol. 1, pp. 375–384. ICSE, Cape Town, South Africa (2010)

Bachelet, Rémi. "Gestion de projet." cour de l'école Centrale de Lille. Disponible [En ligne]: http http://gestiondeprojet.pm/. 2013

Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading, MA (1999)

Balog K. and M. de Rijke. Finding experts and their details in e-mail corpora. In WWW 2006, 2006.

Baron, Naomi S. "Letters by phone or speech by other means: The linguistics of email." Language & Communication 18.2 (1998): 133-170.

Baron 2000 Baron, Naomi S. "Alphabet to email." How Written English Evolved and Where It's Heading (2000).

Baron, N. S., "Why Email Looks Like Speech" in New Media Language, Aitchison, J. and Lewis, D. (ed.), Routledge, London, 2003.

Beck, Kent. Extreme programming explained: embrace change. addison-wesley professional, 2000.

Bekhti S., "DypKM : Un processus dynamique de définition et de réutilisation des mémoires des projets", Thèse de Doctorat, Université de Technologie de Troyes, décembre 2003.

Bekhti, S., Matta, N., 2003. Project memory: an approach of modelling and reusing the context and the de design rationale. In: Proceedings of IJCAI'03 (International Joint of Conferences of Artificial Intelligence), Workshop on Knowledge Management and Organisational Memory, Accapulco.

Bickel, Steffen, Peter Haider, and Tobias Scheffer. "Predicting sentences using n-gram language models." Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2005.

Blake, Catherine, and Wanda Pratt. "Better rules, fewer features: a semantic approach to selecting features from text." Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on. IEEE, 2001.

Blanco, Roi, and Christina Lioma. "Random walk term weighting for information retrieval." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

Blum Kulka, Shoshana, Juliane House, and Gabriele Kasper (eds.) 1989. Cross-cultural pragmatics: Requests and apologies. Norwood : Ablex Publishing.

Bodner, G. M. (1991). Toward a unifying theory. In M. U. Smith and V. L. Patel (Eds), Toward a unifying theory of problem solving (pp 21-23). Lawrence Erblaum Associates: Hillsdale, NJ.

Breuker, J., & Van de Velde, W. (Eds.). (1994). CommonKADS library for expertise modelling: reusable problem solving components (Vol. 21). IOS press.

Brin, Sergey, and Lawrence Page. "The anatomy of a large-scale hypertextual web search engine, 1998." Proceedings of the Seventh World Wide Web Conference. 2007.

Brown, Penelope, and Stephen C. Levinson. "Universals in language usage: Politeness phenomena." Questions and politeness: Strategies in social interaction. Cambridge University Press, 1978. 56-311.

Brown, Penelope, and Stephen C. Levinson. Politeness: Some universals in language usage. Vol. 4. Cambridge university press, 1987.

Brown, Peter F., et al. "A statistical approach to machine translation."Computational linguistics 16.2 (1990): 79-85.

Buckingham Shum, (1996) “Representing Hard-to-Formalise, Contextualised, Muitidisciplinary, Organisational Knowledge” the Workshop on Knowledge Media for Improving Organisational Expertise, Ist International Conference on Practical Aspects of Knowledge Management, Basel, Switzerland, 30-31 October 1996

C

Capobianco, G., De Lucia, A., Oliveto, R., Panichella, A., Panichella, S.: On the role of the nouns in IR-based traceability recovery. In: Proceedings of 17th IEEE International Conference on Program Comprehension. Vancouver, British Columbia, Canada (2009a)

Carvalho, Vitor R., and William W. Cohen. "Improving email speech acts analysis via n-gram selection." Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech. Association for Computational Linguistics, 2006.

Carvalho, Vitor R., and William W. Cohen. "On the collective classification of email speech acts." Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005.

Champin, Pierre-Antoine, Yannick Prié, and Alain Mille. "Musette: Modelling uses and tasks for tracing experience." ICCBR. Vol. 3. 2003.

Charlet Jean. L'ingénierie des connaissances, entre science de l'information et science de la gestion. Working Paper sic00000805, CNRS, France, 2004. http://archivesic.ccsd.cnrs.fr.

Cheung, K., Kwok, J.T., Law, M.H., and Tsui, K., “Mining customer product ratings for personalized marketing”, Decision Support Systems, 35, 2003, 231-243.

Chomsky, Noam. Language and mind. New York: Harcourt Brace Jovanovich, 1972.

Cleland-Huang J. et al. (eds.), Software and Systems Traceability, 3 DOI 10.1007/978-1-4471-2239-5_1, C _ Springer-Verlag London Limited 2012

Cohen, W. W., Carvalho, V. R., and Mitchell, T. M. , “Learning to Classify Email into ‘Speech Acts,’” Proceedings of Empirical Methods in Natural Language Processing (EMNLP2004), 2004.

Cohen, W. W., Carvalho, V. R., and Mitchell, T. M. ,(2004) “Learning to Classify Email into ‘Speech Acts,’” Proceedings of Empirical Methods in Natural Language Processing (EMNLP2004), 2004.

Cohen, William, (1996), “Learning rules that classify e-mail”, AAAI Spring Symposium on Machine Learning in Information Access.

Collobert, Ronan, et al. "Natural language processing (almost) from scratch." The Journal of Machine Learning Research 12 (2011): 2493-2537.

Collobert, Weston 2008) in Proceedings of the 25 th International Conference on Machine Learning, Helsinki, Finland, 2008.

Collot, M., and Belmore, N., “Electronic Language: A New Variety of English,” Herring, S. C Computer-Mediated Communication,. (ed.), John Benjamins, Amsterdam, 1996.

Conklin, E. J., & Weil, W., 1997. Wicked Problems: Naming the pain in organizations (White Paper): Group Decision Support Systems.

Conklin, J., Begeman, M., 1988. gIBIS: a hypertext tool for exploratory policy discussion. In: Proceedings ACM Conference on ComputerSupported Cooperative Work. pp. 140–152

Conklin, Jeff. Dialogue mapping: Building shared understanding of wicked problems. John Wiley & Sons, Inc., 2005.

Coornaert Émile, Les Compagnonnages en France, du Moyen Âge à nos jours, Les Éditions ouvrières, 1966

Cordier, A., Mascret, B., & Mille, A. (2010, July). Dynamic case based reasoning for contextual reuse of experience. In Provenance-Awareness in Case-Based Reasoning Workshop. ICCBR (pp. 69-78).

Core, Mark G., and James Allen. "Coding dialogs with the DAMSL annotation scheme." AAAI fall symposium on communicative action in humans and machines. 1997.

Cormack, Gordon V. "Email spam filtering: A systematic review."Foundations and Trends in Information Retrieval 1.4 (2007): 335-455.

Corston-Oliver, Simon, et al. "Task-focused summarization of email." ACL-04 Workshop: Text Summarization Branches Out. 2004.

Crawford, Vincent P., and Joel Sobel. "Strategic information transmission."Econometrica: Journal of the Econometric Society (1982): 1431-1451.

Cristianini N., J. Shawe-Taylor, An Introduction to Support Vector Machine and other Kernel-based Learning Methods, Cambridge, 2000.

Crystal, D., Language and the Internet. Cambridge University Press, Cambridge, UK, 2001.

Cselle, Gabor, Keno Albrecht, and Rogert Wattenhofer. "BuzzTrack: topic detection and tracking in email." Proceedings of the 12th international conference on Intelligent user interfaces. ACM, 2007.

Culotta A., R. Bekkerman, and A. Mccallum. Extracting social networks and contact information from email and the web. In CEAS-1, 2004.

Culpeper, Jonathan, Dawn Archer, and Matthew Davies. "Pragmatic annotation." (2008): 613-642.

Cunningham, Hamish, Robert J. Gaizauskas, and Yorick Wilks. A General Architecture for Text Engineering (GATE): A New Approach to Language Engineering R & D. University of Sheffield, Department of Computer Science, 1995.

D

Dabbish, L., Kraut, R., Fussell, S., Kiesler, S: Understanding email use: predicting action on a message. SIGCHI conference on Human factors in computing systems (2005).

Dalli, Angelo, Yunqing Xia, and Yorick Wilks. "Fasil email summarisation system." Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004.

Daniel Jurafsky and Martin, James H.,"Speech and language processing." International Edition (2000).

Davenport, Thomas H., and Laurence Prusak. Working knowledge: How organizations manage what they know. Harvard Business Press, 1998.

Davis, A.M.: Just Enough Requirements Management: Where Software Developing Meets Marketing, p. 6. Dorset House Publishing, New York, NY (2005)

De Felice, Rachele, and Paul Deane. "Identifying speech acts in emails: Toward automated scoring of the TOEIC® email task." Princeton: ETS Research Report Series 2012.2 (2012): i-62).

De Felice, Rachele, et al. "A classification scheme for annotating speech acts in a business email corpus." ICAME Journal 37 (2013): 71-105.

De Jong T., & Ferguson-Hessler M. G. M. (1986). Cognitive structures of good and novice problem solvers in physics. Journal of Educational Psychology, 78, 279-288

De Lucia, A., Marcus, A., Oliveto, R., & Poshyvanyk, D. (2012). Information retrieval methods for automated traceability recovery. In Software and systems traceability (pp. 71-98). Springer London.

De Vel, Olivier. "Mining e-mail authorship." Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD'2000). 2000.

Deerwester, Scott, et al. "Indexing by latent semantic analysis." Journal of the American society for information science 41.6 (1990): 391.

Diehl C. P., L. Getoor, and G. Namata. Name reference resolution in organizational email archives. In SIAM Int. Conf. Data Mining 2006, pages 20–22, 2006.

Dieng-Kuntz R., Corby O., Gandon F., Giboin A., Golebiowska J., Matta N., Ribière M., Méthodes et outils pour la gestion des connaissances. 2e edition. Dunod éditeur. 2002.

Diesner, Jana, and Kathleen M. Carley. "Exploration of communication networks from the enron email corpus." SIAM International Conference on Data Mining: Workshop on Link Analysis, Counterterrorism and Security, Newport Beach, CA. 2005.

Dredze, M., Lau, TA, Kushmerick, N.: Automatically classifying emails into activities. Intelligent User Interfaces, 2006, pp :70-77

Drucker, Harris, Donghui Wu, and Vladimir N. Vapnik. "Support vector machines for spam categorization." Neural Networks, IEEE Transactions on 10.5 (1999): 1048-1054.

Dubrovsky, Vitaly J., Sara Kiesler, and Beheruz N. Sethna. "The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups." Human-computer interaction 6.2 (1991): 119-146.

Dumais, S.T.: Improving the retrieval of information from external sources. Behav. Res. Meth. Instrum. Comput. 23, 229–236 (1991)

E

F

Feldman, Ronen, Ido Dagan, and Haym Hirsh. "Mining text using keyword distributions." Journal of Intelligent Information Systems 10.3 (1998): 281-300.

Feng, D., Shaw, E., Kim, J., Hovy, E.: Learning to detect conversation focus of threaded discussions. Human Language Technology Conference, New York (2006).

Feng, Donghui, et al. "Learning to detect conversation focus of threaded discussions." Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006.

Fette, Ian, Norman Sadeh, and Anthony Tomasic. "Learning to detect phishing emails." Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

Fisher, D., Moody, P. 2001. Studies of Automated Collection of Email Records. University of California, Irvine, Technical Report UCI-ISR-02-4.

Freund, Yoav, and Robert E. Schapire. "Large margin classification using the perceptron algorithm." Machine learning 37.3 (1999): 277-296.

G

Gabrilovich, Evgeniy, and Shaul Markovitch. "Feature generation for text categorization using world knowledge." IJCAI. Vol. 5. 2005.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening ontologies with DOLCE. In Knowledge engineering and knowledge management: Ontologies and the semantic Web (pp. 166-181). Springer Berlin Heidelberg

Go, Alec, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." CS224N Project Report, Stanford 1 (2009): 12.

Goldstein et al, 2006. Jade, and Roberta Evans Sabin. « Using speech acts to categorize email and identify email genres » 2006

Goldstein, J., Sabin, R.E.: Using Speech Acts to Categorize Email and Identify Email Genres. System Sciences, HICSS2006 (2006)

Gomez, Juan Carlos, Erik Boiy, and Marie-Francine Moens. "Highly discriminative statistical features for email classification." Knowledge and information systems 31.1 (2012): 23-53.

Gonçalves, João Carlos de AR, Flavia Maria Santoro, and Fernanda Araujo Baiao. "Business process mining from group stories." Computer Supported Cooperative Work in Design, 2009. CSCWD 2009. 13th International Conference on. IEEE, 2009.

Gotel, O., Cleland-Huang, J., Hayes, J. H., Zisman, A., Egyed, A., Grünbacher, P., Mäder, P. (2012). Traceability fundamentals. In Software and Systems Traceability (pp. 3-22). Springer London.

Gotel, O., Finkelstein, A.: An analysis of the requirements traceability problem. In: Proceedings of the 1st IEEE International Conference on Requirements Engineering, Colorado Springs, CO, USA, 18–22 Apr, 1994, pp. 94–101.

Gray, P. H., 2001. A problem-solving perspective on knowledge management practice. In Decision Support System 31, 1 (May 2001), p87-102.

Gruber, T., Russell, D., 1996. Generative design rationale: beyond the record and replay paradigm. In: Moran, T., Carroll, J. (Eds.), Design Rationale: Concepts, Techniques and Use. Lawrence Erlbaum Associates, pp. 323–350 (Chapter 11).

Guzella, Thiago S., and Walmir M. Caminhas. "A review of machine learning approaches to spam filtering." Expert Systems with Applications 36.7 (2009): 10206-10222.

H

Hardin, L. E., 2002. Problem Solving Concepts and Theories. In Journal of Veterinary Medical Education, 30(3), pp. 227-230.

Harzallah, M., Leclère, M., & Trichet, F., 2002. CommOnCV: modelling the competencies underlying a curriculum vitae. In Proceedings of the 14th international conference on Software engineering and knowledge engineering pp. 65-71.

Hassell, Lewis, and Margaret Christensen. "Indirect speech acts and their use in three channels of communication." Proceedings of the First International Workshop on Communication Modeling-The Language/Action Perspective, Tilburg, The Netherlands. 1996.

Hatcher, Erik, and Otis Gospodnetic. "Lucene in action." (2004).

Hatchuel, Armand, Benoit Weil, and Ministere de la Recherche et de la Technologie (MRT), 75-Paris (France); Ecole Nationale Superieure des Mines, 75-Paris (France);. "The expert and the system." (1995).

Hatchuel, Armand, Pascal Le Masson, and Benoît Weil. "De la gestion des connaissances aux organisations orientées conception." Revue internationale des sciences sociales 1 (2002): 29-42.

Haumer, P., Pohl, K., Weidenhaupt, K., Jarke, M., 1999. Improving reviews by extended traceability. In: Proceedings 32nd Hawaii International Conference on System Sciences.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. (?)

Helfman, Jonathan Isaac, and Charles Lee Isbell. "Ishmail: Immediate identification of important information." AT&T Labs. 1995.

Herbsleb, J.D., Kuwana, E., 1993. Preserving knowledge in design projects: what designers need to know? In: Proceedings of the Conference on Human Factors in Computing Systems, April 24–29. ACM Press, New York, pp. 7–14.

Herring (2002b) Dans « Computer-mediated communication: linguistic, social, and cross-cultural perspectives » Pragmatics & Beyond new series

Herring, Susan C. "Computer-mediated communication on the internet." Annual review of information science and technology 36.1 (2002): 109-168.

Herring, Susan C., ed. Computer-mediated communication: Linguistic, social, and cross-cultural perspectives. Vol. 39. John Benjamins Publishing, 1996.

Herring, Susan C., et al. "An approach to researching online behavior." Designing for virtual communities in the service of learning 338 (2004).

I

Ide, N., and Veronis, J., “Introduction to the special issue on word sense disambiguation: the state of the art”, Comput. Linguist., 24 (1), 1998, 2-40.

Ingram, C., & Riddle, S. (2012). Cost-benefits of traceability. In Software and Systems Traceability (pp. 23-42). Springer London.

Ivanovic, Edward. "Dialogue act tagging for instant messaging chat sessions." Proceedings of the ACL Student Research Workshop. Association for Computational Linguistics, 2005.

J

Jose, P.: Sequentiality of speech acts in conversational structure. In: Journal of Psycholinguistic Research. Vol.17, Number 1, pp. 65—88. (1988)

K

Kalia, Anup, et al. Identifying business tasks and commitments from email and chat conversations. Technical Report HPL-2013-4, HP Laboratories, 2013.

Kambhatla, Nanda. "Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations." Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, 2004.

Kantrowitz, Mark, Behrang Mohit, and Vibhu Mittal. "Stemming and its effects on TFIDF ranking (poster session)." Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2000.

Kerbrat-Orecchioni, Catherine, and Henri Mitterand. "Les actes de langage dans le discours(théorie et fonctionnement)." (2008).

Khosravi, Hamid, and Yorick Wilks. "Routing email automatically by purpose not topic." Natural Language Engineering 5.03 (1999): 237-250.

Khoussainov, R., Kushmerick, N.: Email task management: An iterative relational learning approach. CEAS'2005 (2005).

Kiesler, Sara, Jane Siegel, and Timothy W. McGuire. "Social psychological aspects of computer-mediated communication." American psychologist 39.10 (1984): 1123.

Kiesler, Sara. "Hidden messages in computer networks." Harvard Business Review (1986).

Kirsner, Kim, et al. Implicit and explicit mental processes. Psychology Press, 2013 pp 333.

Knethen, A.v.: A Trace Model for System Requirements Changes on Embedded Systems. Proc. of 4th International Workshop on Principles of SW Evolution; Sept. 2001.

Kolodner, Janet. 1993. Case-Based Reasoning Morgan Kaufmann Publishers Inc. San Mateo, CA

Kunz, W., Rittel, H., 1970. Issues as elements of information systemsCenter for Planning and Development Research. University of California, Berkeley

L

Laclavík M., Kvassay M., Dlugolinský Š., Hluchý L.: Use of Email Social Networks for Enterprise Benefit; in: IWCSN 2010, IEEE/WIC/ACM WI-IAT, 2010, pp 67-70

Laclavík, Michal, et al. "Email analysis and information extraction for enterprise benefit." Computing and informatics 30.1 (2012): 57-87.

Laclavík, Michal, et al. "Email social network extraction and search."Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03. IEEE Computer Society, 2011.

Laclavik, Michal, et al. "Ontea: Platform for pattern based automated semantic annotation." Computing and Informatics 28.4 (2012): 555-579.

Lampert 2009a, Andrew, Robert Dale, and Cécile Paris. "Segmenting email message text into zones." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. Association for Computational Linguistics, 2009.

Lampert, Andrew, Cécile Paris, and Robert Dale. "Can requests-for-action and commitments-to-act be reliably identified in email messages."Proceedings of the 12th Australasian Document Computing Symposium. 2007.

Lampert, Andrew, Robert Dale, and Cécile Paris. "Classifying speech acts using verbal response modes." Proceedings of the Australasian Language Technology Workshop. Vol. 2. No. 3. 2006.

Lampert, Andrew, Robert Dale, and Cecile Paris. "Detecting emails containing requests for action." Human Language Technologies: The 2010 Annual Conference of the North

American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010.

Lane, Derek R. "Computer-mediated communication in the classroom: Asset or liability." Workshop presented at the Interconnect '94 Teaching, Learning & Technology Conference October 14, 1994.

Larsen, Kai R., et al. "Analyzing unstructured text data: Using latent categorization to identify intellectual communities in information systems." Decision Support Systems 45.4 (2008): 884-896.

Leuski, Anton. "Email is a stage: discovering people roles from email archives." Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004.Khoussainov, R., Kushmerick, N.: Email task management: An iterative relational learning approach. CEAS'2005 (2005).

Li, H. , Shen, D., Zhang, B. , Chen, A. , Yang, Q.: Adding semantics to email clustering. In:Proceedings of the 6th IEEE International Conference on Data Mining, Hong Kong,China, 2006, pp:938–942.

Li, Wentian. "Random texts exhibit Zipf's-law-like word frequency distribution." Information Theory, IEEE Transactions on 38.6 (1992): 1842-1845.

Licoppe, C., « Logiques d'innovation, multiactivité et zapping au travail », Hermès, La Revue, n°50, 171-178.

M

MacLean, Allan, Richard M. Young, and Thomas P. Moran. "Design rationale: the argument behind the artifact." ACM SIGCHI Bulletin. Vol. 20. No. SI. ACM, 1989.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Symposium on Math, Statistics, and Probability (pp. 281–297). Berkeley, CA: University of California Press.

Malvache, P. and Prieur, P. (1993). Mastering corporate experience with the Rex method. In J. P. Barthès ed., Proc. of ISMICK'93, Compiègne, October, pp.33-41.

Mandelbrot, Benoit B. "7he Fractal Geometry of Nature." (1982): 35.

Manning C. , P. Raghavan, and H. Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.

Marcoccia, M. (2004a) : La citation automatique dans les messageries électroniques, In J.-M. Lopez-Muñoz,S. Marnette, & L. Rosier (eds). Le Discours rapporté dans tous ses états, Paris, L'Harmattan : 467-478.

Marcoccia, Michel (2004). "La communication écrite médiatisée par ordinateur: faire du face à face avec de l'écrit." Journée d'étude sur le traitement automatique des nouvelles formes de communication écrite 5 (2004).

Marcoccia, Michel. "La communication médiatisée par ordinateur: problèmes de genres et de typologie." Journée d'études: les genres de l'oral (2003): 11.

Martin, James H., and Daniel Jurafsky. "Speech and language processing."International Edition (2000).

Martin, James. Rapid application development. Vol. 8. New York: Macmillan, 1991.

Mataka, L. M., Cobern, W. W., Grunert, M. L., Mutambuki, J., & Akom, G. (2014). The Effect of Using an Explicit General Problem Solving Teaching Approach on Elementary Pre-Service Teachers' Ability to Solve Heat Transfer Problems. Online Submission, 2(3), 164-174.

Matta N. Atifi H. Ducellier G., Daily Knowledge Valuation in Organizations, ISTE-Willey 2016.

Matta N., Ducellier G., How to learn from design project knowledge, International Journal of Knowledge and Learning, Int. J. Knowledge and Learning, Vol. 9, Nos. 1/2, 2014

Matta N., Ribière M., Corby O., Lewkowicz M., Zacklad M., Project Memory in Design, Industrial Knowledge Management - A Micro Level Approach, Rajkumar Roy (Eds), Springer-Verlag, 2000

McCandless, Michael, Erik Hatcher, and Otis Gospodnetic. Lucene in Action: Covers Apache Lucene 3.0. Manning Publications Co., 2010.

McDowell, L., Etzioni, O., Halevey, A., Levy, H.:Semantic email.World Wide Web, (2004)

McQuail, Denis. Mass communication. John Wiley & Sons, Inc., 1994.

Metallidou, P. (2009). Pre-service and in-service teachers' metacognitive knowledge about problem-solving strategies. Teaching and Teacher Education, 25, 76-82

Mille Alain Traces Based Reasoning (TBR) Definition, illustration and echoes with story telling (chapter from a book in french ?)

Miller TW (2005). Data and Text Mining. Pearson Education International.

Minkov E., R. C. Wang, and W. W. Cohen. Extracting personal names from emails. In HLT-EMNLP 2005, 2005.

Minkov E., W. W. Cohen, and A. Y. Ng. Contextual search and name disambiguation in email using graphs. In SIGIR '06, pages 27–34, 2006

Morris, Charles William. "Foundations of the Theory of Signs." (1938).

N

Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. "Natural language processing: an introduction." Journal of the American Medical Informatics Association 18.5 (2011): 544-551.

Naur, P., Randell, B. (eds.): Software engineering: Report of a conference sponsored by the NATO Science Committee, Garmisch, Germany, 7–11 October 1968, Brussels, Scientific Affairs Division, NATO (Published 1969)

Newell, Allen, and Herbert Alexander Simon. Human problem solving. Vol. 104. No. 9. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Nonaka I., Takeuchi H.: The knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation. Oxford University Press, 1995

O

Orkin, Jeff, and Deb Roy. "Semi-automated dialogue act classification for situated social agents in games." Agents for games and simulations II. Springer Berlin Heidelberg, 2011. 148-162.

P

Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).

Panckhurst, R. (1998). Analyse linguistique du courrier électronique. In N. Guéguen and L. Tobin (Eds.), Communication, Société et Internet, (pp 47- 60). Paris: L'Harmattan.

Panckhurst, R. (1999). Analyse linguistique assistée par ordinateur du courriel, In J. Anis (Ed.), Internet communication et langue française (pp. 55-70). Paris: Hermès.

Parnas, D., Clements, P., 1985. A rational design process: how and why to fake it? IEEE Transactions on Software Engineering 12, 251–257.

Phan, Xuan-Hieu, Le-Minh Nguyen, and Susumu Horiguchi. "Learning to classify short and sparse text & web with hidden topics from large-scale data collections." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

Pinheiro F.A.C.: Requirements traceability. IN: [PD04], pp.91-113, 2004.

Platon, Œuvres complètes. Théétète, édition de Léon Robin, Belles Lettres (CUF), Paris, 1970Polyani, M. (1966).

Polyani, K. (1944). The great transformation. New York: Rinehart.

Pomian, Joanna. Mémoire d'entreprise: techniques et outils de la gestion du savoir. les Éd. Sapientia, 1996.

Pons-Porrata, Aurora, Rafael Berlanga-Llavori, and José Ruiz-Shulcloper. "Topic discovery based on text mining techniques." Information processing & management 43.3 (2007): 752-768.

Porter, Martin F. "Snowball: A language for stemming algorithms." (2001).

Q

R

Ramesh, B., Edwards, M.: Issues in the development of a requirements traceability model. In: Proceedings of the IEEE International Symposium on Requirements Engineering, San Diego, CA, USA, 4–6 Jan 1993, pp. 256–259.

Ramesh, B., Powers, T., Stubbs, C.: Implementing requirements traceability: A case study.In: Proceedings of the 2nd IEEE International Symposium on Requirements Engineering, pp. 89–95 (1995, March).

Rauscher F, Matta N, Atifi S « Context Aware Knowledge Zoning : Traceability and business Emails » Workshop AI4KM – IJCAI 2015, to be published Springer

Rauscher F., Matta N. and Atifi H. (2015). Hybrid System for Collaborative Knowledge Traceability - An Application to Business Emails. In Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, ISBN 978-989-758-158-8, pages 260-267. DOI: 10.5220/0005640402600267

Reif, F. (1981). Teaching problem solving: A scientific approach. The Physics Teacher, 19, 310-316

Rennie J. , ifile: an application of machine learning to e-mail filtering, in: Proceedings of KDD 2000 Workshop on Text Mining, 2000.

Richard J.F., Les activites mentales, Comprendre, raisonner, trouver des solutions, Armand Colin, Paris, 1990

Rittel, H. W. J. & Webber, M. M. (1973). Dilemmas in General Theory of Planning. Policy Sciences, 4, 155-169

Robinson, Gary. "A statistical approach to the spam problem." Linux journal 2003.107 (2003): 3.

Rocchio J. J. , Document retrieval systems - optimization and evaluation. Ph.D. Thesis, Harvard University, 1966.

Rosas, João, Patrícia Macedo, and Luis M. Camarinha-Matos. "An Organization's Extended (Soft) Competencies Model." Leveraging Knowledge for Innovation in Collaborative Networks. Springer Berlin Heidelberg, 2009. 245-256.

S

Saggion, Horacio, et al. Ontology-based information extraction for business intelligence. Springer Berlin Heidelberg, 2007.

Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613-620.

Scerri, Simon, Siegfried Handschuh, and Brian Davis. "The path towards semantic email: Summary and outlook." Proceedings of EMAIL-08: the AAAI Workshop on Enhanced Messaging. 2008.

Scerri. S., Davis. B., Handschus., S. Improving email conversation efficiency through semantically enhanced email. In 18th international conference on database and expert system applications. Pp 490-494 (2007).

Schapire, Robert E., and Yoram Singer. "Improved boosting algorithms using confidence-rated predictions." Machine learning 37.3 (1999): 297-336.

Schiffrin, Deborah, Deborah Tannen, and Heidi E. Hamilton, eds. The handbook of discourse analysis. John Wiley & Sons, 2008.

Schwaber, Ken. Agile project management with Scrum. Microsoft press, 2004.

Scott, John. Social network analysis. Sage, 2012.

Searle (1969), J.: Speech Acts. Cambridge University Press (1969).

Searle, J.R., 1975. Indirect speech acts (pp. 59-82). http://www.cs.uu.nl/docs/vakken/musy/searle_indirect.pdf.

Sebastiani F (2002). « Machine Learning in Automated Text Categorization."ACM Computing Surveys, 34(1), 1{47. ISSN 0360-0300.

Segal R. B., and Kephart J. O., (1999), "Mailcat: An intelligent assistant for organizing e-mail", Proceedings of the Third International Conference on Autonomous Agents..

Shum, S. B., 1997. Representing hard-to-formalise, contextualised, multidisciplinary, organisational knowledge. In Proceedings of the AAAI Spring Symposium on Artificial Intelligence in Knowledge Management.

Silva, Catarina, and Bemardete Ribeiro. "The importance of stop word removal on recall values in text categorization." Neural Networks, 2003. Proceedings of the International Joint Conference on. Vol. 3. IEEE, 2003.

Singh, M.: A Semantics for Speech Acts. In: Annals of Mathematics and Artificial Intelligence. Vol. 8, Numbers 1- 2, pp. 47—71. (2006)

Smith, Noah A. "Linguistic structure prediction." Synthesis lectures on human language technologies 4.2 (2011): 1-274.

Soares, Diego Carvalho, Flávia Maria Santoro, and Fernanda Araujo Baião. "Discovering collaborative knowledge-intensive processes through e-mail mining." Journal of Network and Computer Applications 36.6 (2013): 1451-1465.

Sowa, J. F., & Dietz, D. (2000). Knowledge representation: logical, philosophical, and computational foundations. Pacific Grove, CA: Brooks Cole Publishing Co.

Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Document. 28, 11–21 (1972)

Sproull, Lee, and Sara Kiesler. "Reducing social context cues: Electronic mail in organizational communication." Management science 32.11 (1986): 1492-1512.

Stiles, William B. Describing talk: A taxonomy of verbal response modes. Newbury Park: Sage Publications, 1992.

Stokes, Nicola, and Joe Carthy. "First story detection using a composite document representation." Proceedings of the first international conference on Human language technology research. Association for Computational Linguistics, 2001.

Stolfo, S. J., Hershkop, S., Hu, C., Li, W., Nimeskern, O., and Wang; K., (2006), “Behavior-based modeling and its application to Email analysis”; ACM Transactions on Internet Technology (TOIT), Volume 6, Number 2, May 2006

Sure, York, Alexander Maedche, and Steffen Staab. "Leveraging Corporate Skill Knowledge-From ProPer to OntoProPer." PAKM. 2000.

Surendran C., John C. Platt, Erin Renshaw. Automatic discovery of personal topics to organize email. In Proc. Conference on Email and Anti-Spam (CEAS)’05, 2005.

T

Tang, A., Jin, Y., & Han, J. (2007). A rationale-based architecture model for design traceability and reasoning. Journal of Systems and Software, 80(6), 918-934.

Tao, Linqing, and David Reinking. "What Research Reveals about Email in Education." (1996).

Tao, Tao, et al. "Language model information retrieval with document expansion." Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Association for Computational Linguistics, 2006.

Tetlock, Paul C., M. A. Y. T. A. L. SAAR-TSECHANSKY, and Sofus Macskassy. "More than words: Quantifying language to measure firms' fundamentals." The Journal of Finance 63.3 (2008): 1437-1467.

Teulier, Régine, and Nathalie Girard. "Modéliser les connaissances pour l'action dans les organisations." L'ingénierie des connaissances (2005): 389

Toulmin, S., 1958. The Uses of Argument. Cambridge University Press.

Tourtier, Paul-André. "Analyse préliminaire des métiers et de leurs interactions." Rapport intermédiaire du projet GENIE, INRIA-Dassault-Aviation (1995).

Tripathi K. and Agrawal M., 2014. Competency Based Management. In Organizational Context: A Literature Review. In Global Journal of Finance and Management. ISSN 0975-6477 Volume 6, Number 4 pp. 349-356.

Trosborg, Anna. Interlanguage pragmatics: Requests, complaints, and apologies. Vol. 7. Walter de Gruyter, 1995.

Turban, B. (2013). Tool-Based Requirement Traceability Between Requirement and Design Artifacts. DOI 10.1007/978-3-8348-2474-5_1, © Springer Fachmedien Wiesbaden 2013)

Turban, B., Kucera, M., Tsakpinis, A., & Wolff, C. (2009). Bridging the requirements to design traceability gap. In Intelligent Technical Systems (pp. 275-288). Springer Netherlands.

Turoff, Murray. "Computer-mediated communication requirements for group support." Journal of Organizational Computing and Electronic Commerce 1.1 (1991): 85-113.

U

V

Vanderveken Daniel, « Illocutionary logic and discourse typology. », Revue internationale de philosophie 2/2001 (n° 216) , p. 243-255 URL : www.cairn.info/revue-internationale-de-philosophie-2001-2-page-243.htm.

Vapnik V., “The nature of statistical learning theory,” Springer-Verlag, New-York, 1995

Vapnik V., “The support vector method of function estimation,” In Nonlinear Modeling: advanced black-box techniques, Suykens J.A.K., Vandewalle J. (Eds.), Kluwer Academic Publishers, Boston, pp.55-85, 1998.

Vergnaud, N., Harzallah, M., and Briand, H, 2004. Modèle de gestion intégrée des compétences et connaissances. EGC pp. 159-170.

Verschueren, Jef. Understanding pragmatics. Oxford University Press, 1999.

W

Walther, Joseph B. "Anticipated ongoing interaction versus channel effects on relational communication in computer-mediated interaction." Human Communication Research 20.4 (1994): 473-501.

Walther, Joseph B. "Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction." Communication research 23.1 (1996): 3-43.

Wasiak, James, et al. "Managing by e-mail: what e-mail can do for engineering project management." Engineering Management, IEEE Transactions on 58.3 (2011): 445-456.

Watzlawick (1972) Une logique de la communication » Paul Watzlawick, Janet. Helmick Beavin, Don D. Jackson éditions du Seuil, 1972 ; 280 p.

Weerkamp, Wouter, Krisztian Balog, and Maarten De Rijke. "Using contextual information to improve search in email archives." Advances in Information Retrieval. Springer Berlin Heidelberg, 2009. 400-411.

Weng, Sung-Shun, and Chih-Kai Liu. "Using text classification and multiple concepts to answer e-mails." Expert Systems with Applications 26.4 (2004): 529-543.

Wiegers, K.E.: Software Requirements: Practical Techniques for Gathering and Managing Requirements Throughout the Product Development Cycle, p. 19. Microsoft Press, Redmond, WA (1999b)

Wilbur, W. John, and Karl Sirotkin. "The automatic identification of stop words." Journal of information science 18.1 (1992): 45-55.

Winograd, Terry, and Fernando Flores. Understanding computers and cognition: A new foundation for design. Intellect Books, 1986.

Winograd, Terry. "A language/action perspective on the design of cooperative work." Human–Computer Interaction 3.1 (1987): 3-30.

X

Y

Yadla, S., Huffman Hayes, J., Dekhtyar, A.: Tracing requirements to defect reports: an application of information retrieval techniques. Innov. Syst. Softw. Eng.: A NASA J. 1(2), 116–124 (2005)

Yang, Yiming, et al. "Topic-conditioned novelty detection." Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002.

Yates, Joanne / Orlikowski, Wanda J., 1993, “Knee-Jerk anti-LOOPism and Other Email Phenomena: Oral, Written, and Electronic Patterns in Computer-mediated Communication”, Technical Report 150, Cambridge MA, Center for Coordination Science, publication en ligne <http://ccs.mit.edu/papers/CCSWP150.html

Yelati, S.; Sangal, R., "Novel Approach for Tagging of Discourse Segments in Help-Desk E-Mails," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on , vol.3, no., pp.369,372, 22-27 Aug. 2011

Z

Zaragoza Hugo 2013 Natural Language Processing for Information Retrieval: an informal overview in 9th European Summer School in Information Retrieval, ESSIR 2013

Zellhofer, S., Collins, M., & Berge, Z. (1998). Why use computer-mediated communication? In Z.L. Berge & M. Collins (Eds.), Wired together: Computer-mediated communication in K-12: Vol. 1: Perspective and instructional design. Cresskill, NJ: Hampton Press.

Zimmermann, A. Context-awareness in user modelling: Requirements analysis for a case-based reasoning application. In Case-Based Reasoning Research and Development, pp. 1064–1064, 2003

Zipf, George Kingsley. "Selected studies of the principle of relative frequency in language." (1932).

Ziv, Oren. "Writing to Work: How Using E-Mail Can Reflect Technological and Organizational Change." Computer-Mediated Communication: Linguistic, social, and cross-cultural perspectives 39 (1996): 243.

François RAUSCHER

Doctorat : Ingénierie Sociotechnique des Connaissances, des Réseaux et du Développement Durable

Année 2016

Gestion des connaissances et communication médiatisée : traçabilité et structuration des messages professionnels

Même si le capital immatériel représente une part de plus en plus importante de la valeur de nos organisations, il n'est pas toujours possible de stocker, tracer ou capturer les connaissances et les expertises, par exemple dans des projets de taille moyenne. Le courrier électronique est encore largement utilisé dans les projets d'entreprise en particulier entre les équipes géographiquement dispersées. Dans cette étude, nous présentons une nouvelle approche pour détecter les zones à l'intérieur de courriels professionnels où des éléments de connaissances sont susceptibles de se trouver. Nous définissons un contexte étendu en tenant compte non seulement du contenu du courrier électronique et de ses métadonnées, mais également des compétences et des rôles des utilisateurs. Egalement l'analyse pragmatique linguistique est mêlée aux techniques usuelles du traitement de langage naturel. Après avoir décrit notre méthode KTR et notre modèle, nous l'appliquons à un corpus réel d'entreprise et évaluons les résultats en fonction des algorithmes d'apprentissage, de filtrage et de recherche.

Mots clés : gestion des connaissances - messageries électroniques - traçabilité - communication dans les organisations - analyse de contenu (communication), logiciels - pragmatique - recherche de l'information.

Knowledge Management and Mediated Communication: Traceability and Structure of Professional Emails

Even if intangible capital represents an increasingly important part of the value of our enterprises, it's not always possible to store, trace or capture knowledge and expertise, for instance in middle sized projects. Email it still widely used in professional projects especially among geographically distributed teams. In this study we present a novel approach to detect zones inside business emails where elements of knowledge are likely to be found. We define an enhanced context taking into account not only the email content and metadata but also the competencies of the users and their roles. Also linguistic pragmatic analysis is added to usual natural language processing techniques. After describing our model and method KTR, we apply it to a real life corpus and evaluate the results based on machine learning, filtering and information retrieval algorithms.

Keywords: knowledge management - electronic mail systems - traceability - communication in organizations - content analysis (communication), software - pragmatics – information retrieval.