Exploratory analysis of the hypertext structure linked to diabetes

Hongyi Shi

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité

Informatique médicale

École Doctorale Pierre Louis de Santé Publique : Épidémiologie et Sciences de l'Information Biomédicale

Présentée par

Mme. Hongyi SHI

Pour obtenir le grade de

DOCTEUR de SORBONNE UNIVERSITÉ

Sujet de la thèse :

Analyse exploratoire de la structure hypertextuelle liée au diabète

Soutenue le 23 juin 2020
Devant le jury composé de :

Mme. Marie-Christine JAULENT	DR / LIMICS, UMRS_1142	Directeur
M. Az-Eddine BENNANI	PU / Université de Techonologie de Compiègne	Rapporteur
M. Amir HAJJAM HASSANI	MC / Université de Technologie de Belfort-Montbéliard	Rapporteur
M. Bernd AMANN	PU / Laboratoire d'Informatique de Paris 6	Examinateur
Mme. Audrey BANEYX	PhD / Sciences Po	Examinateur
M. Nadir AMMOUR	DDS / Global Clinical Innovation Lead, Sanofi R&D	Examinateur
M. Fabien PFAENDER	enseignant-Chercheur / Université de Technologie de Compiègne	Encadrant

Sorbonne Université
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à MON : 01 42 34 68 41
Tél. pour les étudiants de MOO à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

Ce manuscrit s'adresse à toutes les personnes affectées par le diabète et leur entourage.

谨以此文献给所有和糖尿病相知相伴的伙伴，亲友和医护人员们。

This manuscript is for all affected by diabetes and their beloved.

Acknowledgement

This is the last day of 2019, tomorrow will be 2020. I sat in the corner of a café in Paris, watching the people spending their last day of the year.

I still remember the first day of my university for the bachelor degree, it was in 2003. Nothing too much happened to me that year, except I was 18 and have been diagnosed type1 diabetes. Everything looks so fresh, first time drinking, first time smoking, first time going clubbing, first time dying my hair, first time piercing my ears, and of course, first time injecting insulin 4 times per day… Then, time flies, 2020 is coming! I still need to remind myself I am not in a science fiction movie but in the real life.

I never really planned my future, I mean, I even never thought about what I wanted to do in the future before I was diagnosed with diabetes. My father locked me in the room and phoned my mother, telling her I got type 1 diabetes and nobody knew what that was. I thought I would definitely die soon, otherwise why would they locked me in a room and didn't tell me the truth. They told me that I needed to hide my situation, I should not let anybody knows. I tried my best to avoid that anybody sees me injecting the insulin in the toilet even when I lived in a dormitory with others. After 4 years of university life, still nobody knew I have diabetes. My father wanted to save my life by looking for all the possible Chinese medicine which, they say, can cure diabetes. I tried to stop injecting the insulin because I felt ashamed. Because of that, I was sent to the hospital for complications. My mother was praying for me every day and believed in Buddha more and more. They had one common hope, that I shared with them: cure my diabetes and save me from the suffering with all day long injection, monitoring my blood sugar and living my whole life in the fear of worrying for complications and for my future. People say "you don't realize how to live your life until you die once." Being diagnosed as a diabetic is like a rebirth, then I started to think what I would like to achieve in my life before I die.

I thought that I'm not smart enough to get involved in a medical school but I'm still willing to do something for diabetes people. I feel deeply from my heart how painful it is, especially when nobody understands you and take it as abnormal condition in China. We are lacking of education

on diabetes, we are lacking awareness on diabetes, we are even lacking basic medecine resources. After working for International Diabetes Federation as a Chinese ambassador for almost 6 years, I met my co-supervisor Fabien Pfaender in Shanghai, China. He is a French researcher who works in network visualization area in Shanghai. After he discovered my will to help, he encouraged me to study PhD in the visualization diabetes online communities' area. At the same time, he introduced me to my PhD director Marie-Christine Jaulent who works in LIMICS in Paris. They told me I can still make effort to change the situation of diabetes without medicine background, thinking about what is the problem that people with diabetes are really facing and what you can do to solve even only part of the problem. I still remember when Marie-Christine told me that PhD is the work you make for the scientific world; it doesn't matter if it is a small brick or a large one, because your work today is based on others and people after you will continue your job until we achieve some breakthroughs. I really appreciate my two mentors giving me this opportunity to study PhD, to allow me make a little brick in the scientific world and make just a little effort to help people with diabetes. I felt really lucky. How lucky I was to help and to study in the heart of Paris.

Three years PhD life abroad was exciting enough for me to be remembered my whole life, I never expect myself to be diligent as a PhD student. In addition to two tutors who encouraged me with all their heart, taught me, guided me, even edited my two publications word by word, I also want to thank all my colleagues Sonia, Troskah, Naiara, Marion, Isabelle, Gillet, Jacques, Felipe for every moment we shared together and every time they offered me the help. I never felt lonely here because they are my BIG French family.

In addition, I would like to thank Chinese Scholarship Council, without the three years funding for my work, I could not finish my PhD work in Paris, France. Thanks for the last 6 months of INSERM to support me to finish my work.

I would like to thank Médialab of Sciences Po to offer the open workshop for training how to use the tool Hyphe to capture the networking map of diabetes online communities.

In the end, I want to thank my boyfriend Emmanuel Carrere for his supporting and companying. I have to say, PhD is not an easy job. But under his love and trust, I did that. I don't

remember how many times I cried in the dark and told him I can't make that. His patience and faith in me are the most precious thing for me to keep moving forward.

Thanks for my parents and all my friends, people with the same condition as me. If this work has just been a little helpful for you, I see HOPE!

Now I'm studying in diabetes, living with diabetes every day. No more hiding, no more shame, no more fear. I wish everybody with the same condition as me will live in the healthier environment in 2020!

All the best

Mary Shi
Paris, 31 December 2019

List of Publications

1 Shi H, Pfaender F, Jaulent MC, “Mapping the Hyperlink Structure of Diabetes Online Communities”, MedinFo 2019, August 25-30, 2019, Lyon, France.

2 Shi H, Jaulent MC, Pfaender F, “Semantic Interpretation of the map with Diabetes-Related Websites”, The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH 2019), November 4-7, 2019, Coimbra, Portugal.

Résumé

Nous vivons aujourd'hui dans une société où les maladies chroniques telles que le diabète sont de plus en plus répandues. Alors que la médecine moderne a fait des avancées indéniables pour diagnostiquer et prendre en charge le diabète au cours des deux dernières décennies, les patients diabétiques sont au centre de leur dispositif de soin, ayant besoin de soins médicaux continus et importants largement autogérés. Dans l'écosystème du traitement des patients diabétiques, une éducation à la demande et de qualité sur le diabète pour le patient et ses proches joue un rôle essentiel pour l'optimisation du traitement et *in-fine* pour d'amélioration de la qualité de vie des patients. L'accès à des informations correctes et utiles fait partie intégrante de l'éducation sur la maladie. Beaucoup se tournent naturellement vers des ressources en ligne mais ces ressources, souvent gratuites, sont considérablement désorganisées et ne sont pas accessibles à tout le monde en raison des différentes langues et emplacements géographiques voire des moteurs de recherche utilisés pour les chercher. Pour pallier cette désorganisation apparente, l'objectif principal de la recherche réalisée est d'exploiter les nouvelles technologies de l'information pour construire une carte permettant gens pourront naviguer dans les communautés en ligne du diabète pour obtenir des informations et de l'aide notamment sur la vie quotidienne où la médecine n'a que peu de prise. L'hypothèse de ce travail est qu'il existe une communauté en ligne organisée qui relie la plupart des pages créées par des intervenants-contributeurs dans le domaine du diabète et qu'une description sémantique de ces pages Web peut faciliter la recherche d'informations contextualisées et de meilleure qualité que celles offertes par les moteurs de recherche.

Pour atteindre cet objectif, nous proposons une approche qui se décompose en plusieurs étapes successives. La première vise à développer une méthodologie reproductible et durable pour construire et visualiser le réseau des sites Web sur le diabète, en identifiant les liens entre les principaux acteurs du Web de ce domaine. Pour ce faire, nous utilisons deux outils, Hyphe pour la construction du réseau et Gephi pour sa visualisation, deux outils issus de la recherche et complémentaires. Cette méthodologie nous a permis, dans un premier temps, de capturer 430 sites Web avec 6 587 hyperliens et de créer une première version du réseau que nous avons appelé *DiaMap*. Afin de fournir un premier aperçu sur les sous-communautés au sein de la communauté du diabète, nous avons appliqué un algorithme de détection de communautés visant à identifier

des sites qui partagent des similitudes topologiques. Ceci a permis la découverte de 5 classes distinctes identifiées dans *DiaMap* par des couleurs différentes.

Ce premier résultat montre que la communauté en ligne du diabète possède son propre espace au sein du Web et qu'il est organisé en sous localités. Il est remarquable que l'algorithme ne fonctionne que sur des caractéristiques topologiques de graphe qui mettent évidence des communautés qui partagent effectivement un intérêt commun qui transcende leur existence dans un monde numérique puisque ces communautés sont incarnées par leurs participants.

Pour la deuxième étape, nous avons proposé une approche sémantique pour interpréter le réseau *DiaMap*. Cette approche comporte elle-même trois étapes : (1) définir et organiser les balises textuelles décrivant le contenu sémantique des sites Web ; (2) mettre en place un processus d'annotation pour annoter manuellement les 430 sites Web à l'aide des balises; (3) appliquer diverses méthodes d'apprentissage automatique pour prédire la communauté issue de la topologie d'un site Web à partir des annotations. Nos résultats montrent une performance prédictive assez inégale et relativement moyenne, en utilisant des balises pour prédire les clusters de *DiaMap*. Cela renseigne en revanche sur la réalité de la composition des communautés : un mélange de sites Web de différents types qui créent un espace mixte sémantiquement mais localisé. Cela a également montré que la communauté peut dans certains cas avoir une identité forte sémantiquement parlant. L'approche sémantique demeure donc appropriée pour la recherche d'informations de qualité mais ne peut être la seule source de classification.

Enfin, et pour revenir à l'objectif initial d'aide à l'orientation pour les patients diabétiques, nous avons établi un protocole pour déterminer comment ces derniers peuvent tirer parti des balises pour localiser les sites Web plus rapidement et plus précisément. Le protocole compare le résultat de requêtes via des moteurs de recherche à une interrogation directe de *DiaMap*. *DiaMap* présente la visualisation d'informations de type cartographie de sites Web liés au diabète pour proposer une image du diabète dans le monde numérique. Différent des moteurs de recherche traditionnels, *DiaMap* présente l'ensemble de l'espace du Web du diabète et utilise des balises pour identifier les sites Web pertinents. On peut alors utiliser une navigation alternative dans les informations en

ligne sur le diabète et cela pourrait être une alternative pour augmenter les moteurs de recherche actuels dans le cas des maladies chroniques.

Pour aller plus loin, nous souhaitons augmenter la taille du corpus de sites Web pour incorporer dans *DiaMap* plus de sites dans différentes langues. Notre idée est d'utiliser le traitement automatique du langage naturel (TALN) pour apprendre les balises / topics les plus pertinents des sites Web et annoter automatiquement un ensemble de données de plus grande taille.

Ce travail de thèse conclut qu'un espace en ligne sur le diabète existe et que cet espace virtuel est composé de ressources importantes et devrait être connu de tous. Il contribue en outre à la mis en place d'une méthodologie systématique proposant de combiner la visualisation de réseau et la détection communautaire pour faire de l'analyse de réseau dans le diabète. Cette nouvelle approche peut être appliquée à l'analyse de réseaux pour d'autres pathologies.

Contexte

Le diabète est une maladie chronique qui survient soit lorsque le pancréas ne produit pas suffisamment d'insuline, soit lorsque le corps ne peut pas utiliser efficacement l'insuline qu'il produit. Selon un rapport de l'OMS, environ 1,6 million de décès ont été directement causés par le diabète en 2015. Le diabète est également une cause majeure de cécité, d'insuffisance rénale, de crises cardiaques, d'accidents vasculaires cérébraux et d'amputations des membres inférieurs. Le coût économique mondial du diabète en 2014 était estimé à 612 milliards de dollars américains. Au cours des deux dernières années, il y avait 415 millions de personnes atteintes de diabète et ce chiffre pourrait atteindre 642 millions d'ici 2040. Par conséquent, la récente augmentation spectaculaire des cas de diabète soulève plusieurs questions qui doivent être abordées : comment réagissent les personnes après leur diagnostic *Diabète* ? Que feront-ils pour une autogestion à long terme ? Où peuvent-ils trouver des informations pertinentes et est-il facile d'y accéder ? Comment peuvent-ils obtenir un soutien psychologique et une aide communautaire ?

Le diabète peut être traité et ses conséquences pourraient être évitées ou retardées avec un régime alimentaire spécifique, de l'activité physique, des médicaments, un dépistage et un

traitement réguliers des complications. Une liste préliminaire des principales parties prenantes possibles peut être établie : hôpitaux, médecins, laboratoires de recherche, sociétés pharmaceutiques, structures de relais médical et social, associations, ONG liées à la santé (organisations non gouvernementales), patients et leurs familles, publications sur les soins du diabète. Tous ces acteurs clés du diabète jouent leur propre rôle dans le système de soins de santé, mais la nature de leurs relations les uns avec les autres, le cas échéant, reste inexplorée en dehors d'un paradigme centré sur le patient. Notre hypothèse est que tous les acteurs impliqués dans le diabète ont des connexions qui peuvent être révélées sur le World Wide Web (en abrégé WWW ou le Web) comme un monde organisé de communautés plutôt que comme un réseau organisé de manière aléatoire.

Depuis son invention au début des années 1990, le Web est progressivement devenu une plateforme médiatique majeure hébergeant elle-même des milliards de ressources accessibles à partir d'une URL (Uniform Resource Locator) et des dizaines de sous-médias sous la forme de sites Web spécifiques hébergeant les médias sociaux (comme Facebook ou Twitter), des connaissances scientifiques ou tout autre créneau. Le Web est également la source de la recherche scientifique en soi et la structure de réseau d'un environnement hyperlien peut être une source d'information riche pour qui recherche un moyen efficace de le collecter, de l'analyser et de le comprendre.

L'exploration des structures Web est un proxy pour les structures organisationnelles humaines. Ainsi, nous proposons de cartographier la structure des hyperliens des acteurs du diabète en ligne. En fait, le Web contient des aspects cruciaux de l'incorporation des facteurs sociaux : blogs personnels, sites Web institutionnels, médias axés sur la santé, etc. Pour aborder la relation avec les parties prenantes dans le domaine du diabète, nous avons examiné comment les études classiques sur l’exploration des réseaux peuvent être mises à profit dans le contexte du diabète pour répondre aux questions suivantes : Qui est connecté à qui par quels moyens ? Quelles organisations reçoivent de l'aide de quelles organisations ? Quelles ressources ou informations sont publiées sur quelles plateformes ? Quelle est la relation entre les organismes de bienfaisance et les agences gouvernementales ? Comment les individus, les entreprises et les organisations interagissent-ils ensemble ? Quelle est l'écosphère du monde du diabète ? Dans l'ensemble, notre

objectif est de fournir une méthodologie pratique reproductible pour visualiser le réseau de sites Web sur le diabète afin de s'assurer que tous les principaux acteurs du Web sur la santé sont connectés les uns aux autres. Et puis, les questions précédentes peuvent être abordées en étudiant la carte résultante des communautés en ligne du diabète. Un impact corollaire attendu est que cette carte peut aider les diabétiques à accéder à des informations utiles concernant le contexte voisin (telles que des informations sur l'alimentation ou l'activité physique) que les moteurs de recherche n'affichent pas immédiatement lorsque des informations sur le diabète sont recherchées. La représentation visuelle sera utile pour identifier rapidement et efficacement les zones du graphique où des améliorations peuvent être apportées.

État de l'art

Ce manuscrit utilise les propriétés du web comme un espace non euclidien, hébergeant des ressources en réseau parmi lesquelles des pages hyper textuelles regroupées en sites web et qui créent des sous-espaces. Le web lui-même est soutenu par une couche physique appelée Internet permettant la transmission entre un serveur hébergeant une ressource et un consommateur. Il utilise un protocole de communication établi, à savoir HTTP et HTTPS. Internet est incarné par des appareils physiques dans un lieu appartenant à un pays. Par conséquent, le contenu transféré peut être soumis à la censure ou à des modifications arbitraires de la part dudit pays, ce qui affecte la nature même de toute enquête scientifique sur le Web et les communautés qu'il soutient. Nous avons adapté notre méthodologie pour résoudre ce problème. L'étude se concentre sur les ressources Web liées au diabète pour trouver leurs relations et leurs sous-espaces ultérieurs. Une initiative à long terme, le web sémantique, consiste à expliciter le sens des sujets, des concepts dans les ressources web et les relations entre eux directement dans les ressources elles-mêmes. Nous n'avons pas utilisé les propriétés du web sémantique pour ce travail mais nous nous sommes plutôt concentrés sur (1) l'exploration des propriétés topologiques du réseau de ressources comme le ferait un utilisateur moyen et (2) la proposition d'une interprétation sémantique a posteriori de la structure effectivement observée.

Par conséquent, l'analyse de réseau tient une grande place dans la création de cartes. L'analyse de réseau elle-même a diverses racines, théories et concepts dérivés du domaine qui

l'utilise. Tout d'abord, pour vérifier le sous-espace du Web consacré au diabète extrait par l'outil d'exploration Hyphe (*https://hyphe.medialab.sciences-po.fr*), nous l'avons comparé à d'autres réseaux Web supposés similaires, eux-mêmes faisant partie d'une classe de réseaux complexes. Un réseau complexe se caractérise par trois propriétés principales : la distribution des degrés, le coefficient de regroupement et la longueur moyenne du trajet. Nous avons calculé ces propriétés sur notre réseau pour valider son comportement en tant que réseau Web. Deuxièmement, comme notre sous-espace doit représenter une organisation sociale intrinsèque, nous avons utilisé des mesures de centralité de l'analyse des réseaux sociaux ainsi que des détections communautaires. Ces outils nous permettent d'extraire les communautés du réseau elles-mêmes un sous-espace de l'espace du diabète. Troisièmement, un objectif de la présente étude est d'améliorer l'efficacité pour l'internaute moyen de récupérer des informations sur un sujet spécifique lié au diabète et d'obtenir une connaissance contextualisée de la place du sujet dans la communauté du diabète. Pour atteindre cet objectif, nous devions comparer nos résultats avec les résultats prévus des moteurs de recherche tels que Google. Nous avons appliqué l'algorithme de classement des pages très populaire pour examiner le classement prévu de ressources spécifiques dans un réseau construit autour du diabète (ce que le moteur de recherche aurait proposé) et l'avons comparé à nos choix organisés par l'homme sur la base de la carte.

Lorsque l'analyse de réseau permet un examen détaillé des caractéristiques du réseau, la compréhension de sa complexité nécessite un autre paradigme, dans ce cas : la visualisation. Les données et la visualisation de réseau font partie d'une méthodologie pour créer des images puissantes et perspicaces avec les données tout en contrôlant soigneusement ce que l'utilisateur de la visualisation peut découvrir ou conclure en s'appuyant sur sa capacité visuelle. La visualisation des données a trois objectifs principaux : fournir des informations sur une situation complexe pour le chercheur, aider à vérifier la validité des informations et communiquer les découvertes à un public plus large. Pour étudier le réseau du diabète, nous avons pleinement utilisé ces objectifs, en examinant les réseaux avec une riche représentation interactive aidant à formuler des hypothèses jusqu'à fournir une carte du diabète au grand public pour obtenir une représentation organisée mais précise de l'organisation communautaire du diabète en ligne.

Méthodologie

Afin de visualiser le réseau du diabète dans le monde numérique représenté par le Web pour trouver les connexions des principaux acteurs du Web de la santé les uns avec les autres et pour comprendre la nature de leur relation, on peut utiliser un logiciel de visualisation de grands graphiques matures. La visualisation graphique est une approche efficace pour spatialiser les réseaux complexes. L'utilisation d'un affichage visuel permet d'identifier les caractéristiques d'une structure de réseau et de données tout en s'appuyant sur des capacités perceptuelles humaines très efficaces.

Avec cet objectif clé global à l'esprit, nous décomposons la recherche en phases. Le premier objectif est de développer une méthodologie reproductible et durable pour visualiser le réseau diabétique de sites Web, en identifiant la connexion entre les principaux acteurs du Web des sciences de la vie et des soins de santé. La création puis l'analyse d'un réseau de sites Web thématiques est une méthodologie qui a été appliquée à divers contextes, mais pas encore dans le domaine de la santé. Afin de produire une carte de visualisation des communautés en ligne du diabète qui n'existe pas, nous proposons de combiner deux outils existants distincts qui pourraient soutenir la création d'une telle carte dans le contexte du diabète. La carte (nous l'appelons *DiaMap*) fournira des procurations pour générer des informations clés sur la communauté. Pour ce faire, nous utilisons deux outils de pointe, Hyphe et Gephi (*https://gephi.org*), qui sont complémentaires pour créer une méthodologie robuste recueillant les liens des sites Web et les visualisant en détail. Ensuite, pour fournir une première vue des sous-communautés à l'intérieur de *DiaMap,* nous avons appliqué un algorithme de détection de communautés visant à détecter des grappes de nœuds connectés de manière similaire. Il le fait en maximisant la métrique de qualité de modularité dans toutes les partitions possibles. La modularité mesure la différence entre la densité des bords dans la partition et un graphique randomisé avec le même nombre de nœuds et la même distribution de degrés.

Dans la deuxième partie, nous avons proposé une approche sémantique pour interpréter le réseau *DiaMap*, qui s'est divisée en trois étapes : (1) définir et organiser les balises décrivant le contenu sémantique des sites Web ; (2) mettre en place un processus d'annotation pour annoter

manuellement l'ensemble des données à l'aide des balises; (3) d'appliquer diverses méthodes d'apprentissage automatique pour prédire la classe d'appartenance d'un site Web à partir des annotations.

Nous avons utilisé l'analyse thématique inductive qui est l'une des méthodes d'analyse qualitative populaire pour construire des catégories et décrire le contenu des sites Web avec différentes dimensions. Un expert sur le diabète avec 15 ans d'expérience en diabète de type 1 a proposé l'ensemble initial des catégories pour annoter les sites Web en fonction des parties prenantes du diabète, de la langue et du type des sites Web liés au diabète, du type de diabète et des sujets liés au diabète, etc. Ensuite, l'équipe du projet a examiné une partie des sites Web au hasard et a annoté chaque site Web en utilisant cet ensemble de catégories. À partir des propositions, nous avons défini des catégories associées à des jeux de valeurs possibles constituant les balises finales pour le processus d'annotation. L'expert sur le diabète a annoté manuellement l'ensemble des données avec les balises finales.

Pour étudier quelles combinaisons de balises peuvent prédire ou expliquer les classes / communautés / clusters obtenus par l'algorithme de détection de communautés, nous avons utilisé le framework RapidMiner studio (*https://rapidminer.com*) pour appliquer 7 modèles de prédiction les plus répandus aujourd'hui à notre ensemble de données. RapidMiner Studio est un logiciel de conception de flux de travail visuel permettant d'importer, de préparer et de nettoyer les données, puis d'exécuter un algorithme de science des données et d'apprentissage automatique de pointe. Les méthodes de modélisation sont les suivantes : Naive Bayes, modèle linéaire généralisé, apprentissage profond, arbre de décision, forêt aléatoire, arbres boostés par gradient et machine à vecteur de support. À des fins de test, l'ensemble de données a été divisé en un ensemble d'exemples résolus pour la phase d'apprentissage de l'algorithme et un ensemble de test pour déterminer la qualité de la prédiction. Nous avons utilisé 38 balises pour décrire les données et l'identifiant de la classe comme étiquette pour alimenter différents modèles. Nous avons choisi un ensemble de 60% de sites Web aléatoires pour l'apprentissage et 40% pour les tests et répété plusieurs fois la procédure. Ce ratio pour l'apprentissage / les tests pourrait être plus élevé mais avec le risque de sur-adapter les données.

Enfin, nous avons établi un protocole pour déterminer comment les utilisateurs peuvent tirer parti des balises pour localiser les sites Web plus rapidement et plus précisément en comparant les moteurs de recherche à *DiaMap*. Cinq requêtes sont proposées pour demander aux moteurs de recherche de trouver les sites Web qui peuvent répondre correctement à ces requêtes. Dans le même temps, nous exprimons ces requêtes par le biais des balises utilisées pour décrire tous les sites Web dans *DiaMap*. Nous avons alors d'une part exécuté les requêtes pour demander aux moteurs de recherche de rechercher les sites Web appropriés et d'autre part utilisé les balises correspondantes pour identifier les sites Web pertinents dans *DiaMap*. Nous faisons la comparaison entre tous les résultats offerts par les moteurs de recherche et *DiaMap* pour comprendre comment les utilisateurs peuvent obtenir ce qu'ils veulent à un certain moment et à quel point les résultats présentés par la liste des sites Web et le graphique sont satisfaisants.

Outils

Hyphe est un robot d'exploration Web, développé par le département Médialab de Sciences Po à Paris, en France. L'outil permet de construire un corpus Web en explorant des pages Web et en créant des réseaux entre des « entités Web », connectées les unes aux autres à l'aide d'hyperliens. Hyphe a été utilisé avec succès dans de nombreux articles publiés en sciences sociales. A titre d'exemple, Hyphe permet de rassembler méthodiquement des entités web et de visualiser le réseau destiné aux journalistes de données. Une équipe de bibliothécaires et de sociologues a délimité le débat sur le changement climatique sur le web, mesurant notamment la forte présence inattendue de sceptiques climatiques. De plus, les gens ont essayé d'étudier à travers le mur Facebook « Alternative pour l'Allemagne » (AFD) qui est déjà devenu l'un des plus grands forums de droite sur l'Internet germanophone pour voir comment les médias sociaux jouent un rôle crucial pour la stratégie de mobilisation du parti. Pourtant, Hyphe n'a jamais été utilisé dans le domaine de la santé publique pour évaluer la structure des communautés en ligne sur les maladies chroniques. Par ailleurs, même si Hyphe inclut des capacités de visualisation de réseau de base, il atteint rapidement ses limites à mesure que les réseaux se développent au-delà de cent nœuds.

Gephi est un logiciel « open source » pour la visualisation et l'analyse de réseaux. Il utilise un moteur de rendu 3D pour afficher de grands réseaux en temps réel permettant aux chercheurs

d'explorer leur réseau. Gephi intègre une grande variété d'algorithmes, de filtres et de dispositions pour fournir un réglage fin sur l'analyse et l'affichage du réseau. Il peut gérer un grand réseau (c'est-à-dire plus de 20000 nœuds) et, parce qu'il est construit sur un modèle multi-tâches, il tire parti des processeurs multicœurs. De plus, les conceptions des nœuds et des arcs peuvent être personnalisées pour créer une visualisation semblable à une carte à des fins de communication, idéale pour une carte de la communauté du diabète.

Dans ce contexte, nous avons développé une méthodologie pour découvrir l'organisation en ligne des sites Internet sur le diabète. Nous avons exploité et tiré le meilleur parti de ces outils pour collecter des informations à l'aide d'Hyphe, puis visualiser leurs propriétés et leurs liens à l'aide de Gephi pour produire le réseau sur des sites Web sur le diabète et enfin fournir des informations clés sur cette communauté.

Déroulement de la captation

La méthodologie de captation proposée consiste en un processus intégrant les étapes suivantes :

- collecter des sites Web communautaires sur le diabète;
- analyser la structure des sites Web qui en résulte;
- visualiser le réseau communautaire du diabète comme une carte de navigation.

Ces étapes peuvent être décomposées en un processus plus détaillé potentiellement applicable à toute analyse de communauté en ligne (voir figure 1). La méthodologie suit donc les étapes suivantes :

1. rassembler une liste de sites Web de départ ;
2. créer un corpus de pré-exploration ;
3. consolider le corpus pré-crawl ;
4. lancer et exécuter une exploration à grande échelle ;
5. extraire le réseau de sites Web ;
6. visualiser la structure du réseau.

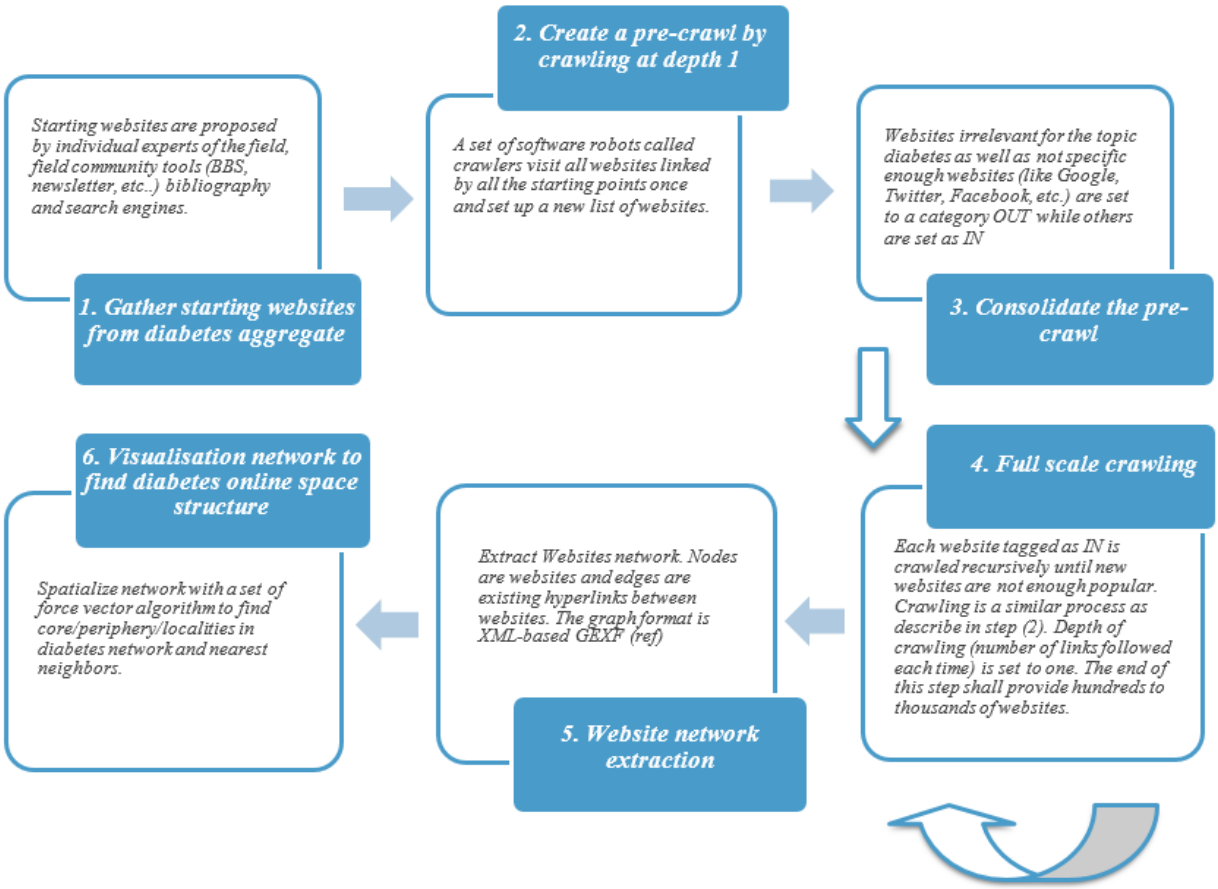


Figure 1 Illustration du flux de travail

Ces six étapes sont basées sur les critères de décision précis à chaque étape.

Collection de sites Web de départ : Afin d'explorer une communauté en ligne composée de sites Web liés à des sujets, nous devons définir des points d'entrée dans la communauté. Ces points d'entrée serviront de porte d'entrée dans la communauté en suivant les liens des points avec un robot soigneusement contrôlé. Comme l'objectif est d'explorer tous les aspects de la communauté, les points d'entrée doivent refléter différents points de vue, opinions ou sujets que la communauté pourrait aborder afin de s'assurer de capturer suffisamment de portes menant aux différentes parties potentielles du réseau final. Pour assurer la diversité des points d'entrée, nous avons sélectionné une variété de sources en commençant par (a) deux ou plusieurs experts du sujet suggérant des sites Web (pour le sujet du diabète, ces experts ont été trouvés chez des personnes atteintes de diabète et travaillant également dans le domaine du diabète depuis plusieurs années);

(b) en utilisant diverses requêtes sur le diabète sur Google pour garantir des résultats différents et (c) en utilisant des suggestions de pages de réseaux sociaux communautaires sur le diabète.

Choix de la profondeur et des critères d'arrêt de l'exploration : Une fois les points de départ définis, Hyphe offre la possibilité d'avoir trois analyses de la profondeur, de 1 clic à chaque hyperlien à 3 clics en fonction de la profondeur d'exploration des sites Web. Comme le réseau Web suit une distribution des pages en loi de puissance, une exploration en profondeur 3 (page de départ – 20 pages à 1 clic – 400 pages à 2 clics – 8000 pages à 3 clics) peut potentiellement rassembler des milliers de sites Web, ce qui rend difficile l'évaluation par un humain de chacun des sites de la communauté du diabète ciblée. Dans cette étude, pour éviter trop de bruit (c'est à dire des sites web qui ne seraient pas en relation avec le diabète), nous avons décidé de ne considérer qu'un niveau de profondeur de 1, répété plusieurs fois si nécessaire lorsque les sites ciblés proposent une contribution importante au sujet. De plus, les critères pour arrêter l'exploration de nouvelles entités Web (unité éditoriale homogène qui peut être un site ou une page ou un sous-site, etc) dépendent de la quantité de sites Web liés au diabète qui apparaissent dans les résultats finaux. Si aucune nouvelle entité liée au diabète n'apparaît après l'exploration, nous considérons avoir atteint une frontière locale dans la communauté du diabète. Il faut alors poursuivre avec les entités Web restantes jusqu'à ce qu'aucune entité ne soit laissée non explorée et que toutes les frontières de sujet soient découvertes. Enfin, compte tenu du grand nombre d'entités web dans le web du diabète, nous n'avons pas considéré dans cette première étude les entités mentionnées par une seule autre entité. Autrement dit une entité était considérée comme valide lorsqu'au moins trois autres entités en faisait mention. Il s'agit d'une propriété importante des réseaux Web fortement connectés où une entité remarquable doit être reliée par plusieurs voisins. Nous avons considéré un minimum de 3 hyperliens vers une entité que nous qualifions de « populaire » pour qu'elle soit éligible à une exploration supplémentaire de profondeur 1.

Processus de nettoyage pour alimenter chaque itération de l'étape 4 avec une base de données IN : Pour nettoyer la base de données après chaque itération, nous avons utilisé deux façons de prendre une décision et de classer les sites Web comme : IN, OUT ou UNDECIDED. En effet, dans cette étude, nous avons classé le contenu des entités Web strictement liées au diabète comme IN; lorsque rien n'indique un contenu potentiel lié au diabète, nous l'avons classé comme

OUT; enfin, les sites Web ambigus qui mentionnent le diabète entre autres sujets ont été classés comme UNDECIDED en attendant une analyse plus approfondie (de leurs liens par exemples) pour lever le voile sur leur classification. La première heuristique pour classer les entités est de sélectionner automatiquement dans la base de données les URLs contenant le mot « diabète » (ou son équivalent en anglais diabetes) et de les définir par défaut comme IN. Ensuite la seconde heuristique de classement des entités est de sélectionner manuellement la classe après un examen du contenu de chaque site Web. Cela se fait en ouvrant son URL dans un navigateur et à le faire évaluer par un expert du diabète comme étant lié ou non au sujet. Après la classification, nous pouvons filtrer la base de données pour ne conserver que les entités IN comme points de départs et procéder à une nouvelle itération. On garde les UNDECIDED comme potentiel candidat à ce stade.

Choix du modèle final à l'étape 6 (Gephi) : Une fois que le « pré-crawl » a été créé en explorant les sites Web de départ à la profondeur 1 et en les nettoyant, nous avons à nouveau appliqué le processus itératif « crawl & clean » plusieurs fois pour obtenir un corpus exploré à grande échelle (étape 4). Ainsi chaque entité Web considérée comme IN est explorée de manière récursive jusqu'à ce qu'aucun nouveau site Web ne soit suffisamment populaire, cité plus de 3 fois pour être inclus ou hors sujet (OUT ou UNDECIDED). La fin de cette étape fournit des centaines de sites Web IN et leurs hyperliens. Nous avons ensuite exporté le réseau de entités Web sous forme de graph au format GEXF (format de graphe basé sur XML). Dans le graphe final, les nœuds sont des entités web assimilés ici dans 100% du temps à des sites Web et les arêtes sont des hyperliens existants entre les sites Web (étape 5). Enfin, nous avons importé le graphe dans Gephi pour en visualiser la structure et révéler les communautés en ligne du diabète (étape 6). Pour ce faire, après avoir importé le réseau, nous avons utilisé une mise en espace du graphe appelé « Force Atlas 2 » (modèle où les nœuds se repoussent mais les liens les attirent) particulièrement adaptée à la spatialiser des graphes de sites Web.

Principales conclusions

Au final, nous avons réussi à capturer 430 sites Web reliés par 6 587 hyperliens du cœur du graphe du diabète sur internet : *DiaMap*. Une carte en couleur est associée à la classe des 5

communautés découvertes au travers d'un algorithme de détection de communautés dans les graphes, ce qui propose la figure 2.

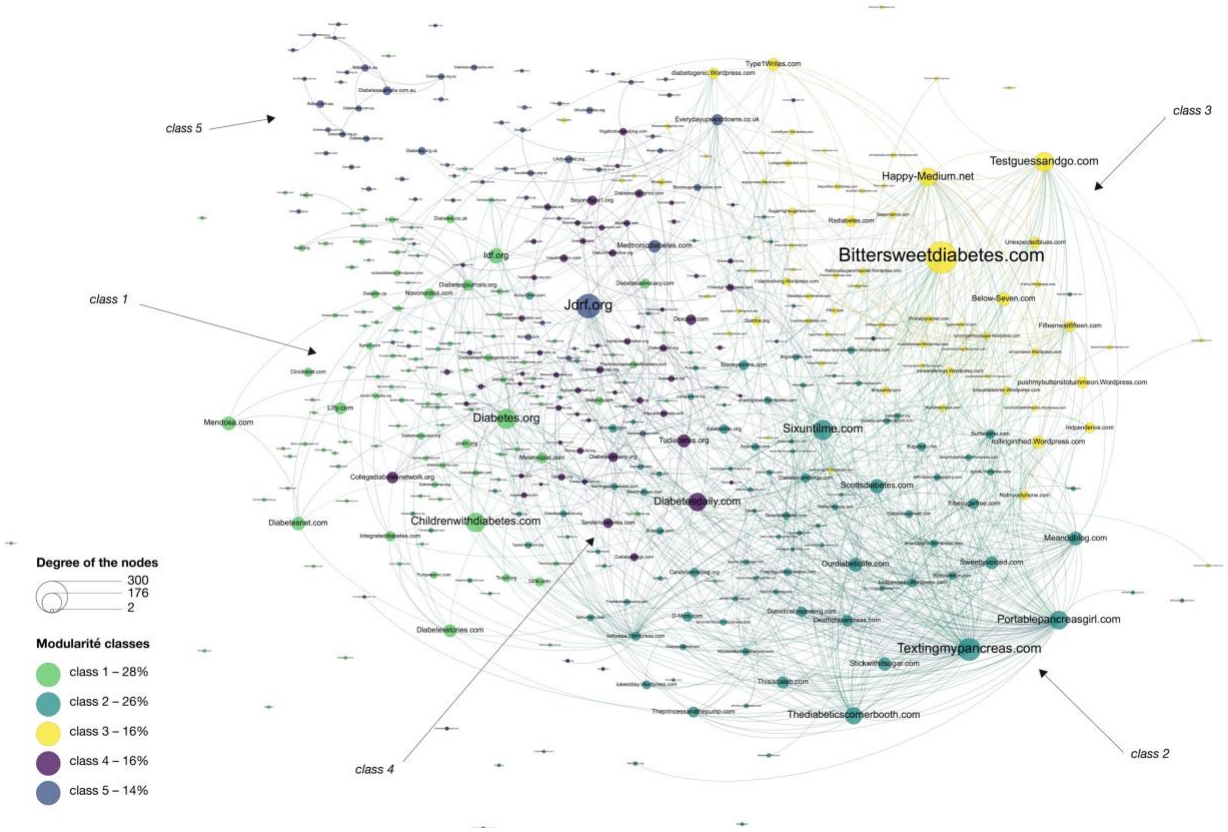


Figure 2 Structure hyperlien du diabète avec disposition de l'atlas de force 2 et communautés

Chaque site Web est associé à la classe à laquelle il appartient. Ici, nous utilisons la classe 1, la classe 2, la classe 3, la classe 4 et la classe 5 comme notations arbitraires pour distinguer les classes.

Analyse Sémantique du corpus

Une fois le corpus constitué, conformément à la méthodologie nous avons défini et organisé les balises décrivant le contenu sémantique des sites Web. Ce travail a été mené dans le but de catégoriser les sites Web en aval. Il est le fruit d'une concertation nourrie par un expert du diabète

en conjonction avec deux chercheurs expérimentés dans le domaine des ontologies et du web. Les balises finales ne sont pas uniquement celles du domaine que l'on pourrait trouver dans une ontologie de domaine classique mais dépasse ce cadre pour qualifier également les types de médias rencontrés, leur structure, leur public ou leurs auteurs, etc. La liste finale des balises est un ensemble hétérogène de catégorie décrivant donc tour-à-tour ces différents attributs et les constituer en ontologie relèverait d'un travail très important qui n'était la direction souhaitée ici

Les balises sont réparties dans six catégories : Statut des parties prenantes, Langue des sites Web, Type de sites Web, Organisations, Type de diabète et Thèmes liés au diabète. Les 6 catégories fournissant 38 valeurs sont présentées dans le tableau 1. Pour certaines d'entre elles, les valeurs s'excluent mutuellement en tant que statut, langue, type de sites Web, organisations et pour les autres, plusieurs valeurs sont autorisées en tant que type de diabète et sujets liés au diabète.

Tableau 1. Résultat de 6 catégories avec 38 valeurs pour le balisage de 430 sites Web liés au diabète.

Status of Stakeholders	**Language of Websites**	**Type of Websites**	**Organizations**	**Type of Diabetes**	**Diabetes-Related Topics**	
Non-Profit	English	Portal	Individual	Type 1	Prevention	
Profit	Multilingual	Information	Association	Type 2	Treatment	
		Blog	Society	Gestational	Self-management	
		Forum	Federation	Pre-diabetes	Advocacy	
		E-commerce	Charity		Complications	
		Click-to-donate	Company		Psychological Support	
			Program		Accessories	
			Conference		Sport	
			Hospital		Diabulimia	
			Clinic			
			Pharmacy			
			Laboratory			
			Consulting			
			Media			
			Online Community			**Total**
2	2	6	15	4	9	**38**
Maximum number of possible tags from each category to annotate one web-site						
1	1	1	1	4	9	**17**

Les balises ont été appliqués par un expert du diabète à chaque entité Web dans Hyphe en les visitant une à une.

À cette étape nous avons donc 430 sites Web, répartis dans des communautés topologiques. On souhaite à présent trouver si ces communautés purement topologiques liées à la structure du graphe sont traversés par une unité sémantique. Chaque entité possède donc une communauté identifiée et un ensemble de balises qui en décrive le contenu et le contenant. On va alors chercher à prédire la communauté à partir des balises. Si l'on y parvient, alors on pourra évaluer l'unité sémantique de chaque communauté topologique, avérer le lien entre la sémantique et la structure de graphe et proposer un guide intelligible dans le corpus en ligne du diabète.

Apprentissage automatique

Sept modèles classiques d'apprentissage automatique ont été appliqués aux données obtenues par le processus d'annotation pour prédire les communautés en fonction des balises. Les performances correspondent au nombre de fois où la machine peut prédire correctement la classe du site Web à partir du modèle de balises. **La précision est définie comme le nombre de vrais positifs sur le nombre de vrais positifs plus le nombre de faux positifs**. Autrement dit combien de sites sont effectivement dans la classe prédite divisé par le total de site prédits pour cette classe. **Le rappel (*recall*) est défini comme le nombre de vrais positifs par rapport au nombre de vrais positifs plus le nombre de faux négatifs**. Autrement dit combien de sites sont effectivement dans la classe prédite divisé par le total de sites qui devraient y figurer.

Un système avec un rappel élevé mais une faible précision renvoie de nombreux résultats, mais la plupart de ses étiquettes prédites sont incorrectes par rapport aux étiquettes d'apprentissage. Un système avec une haute précision mais un faible rappel est tout le contraire, retournant très peu de résultats, mais la plupart de ses étiquettes prédites sont correctes par rapport aux étiquettes d'apprentissage. Un système idéal avec une haute précision et un rappel élevé retournera de nombreux résultats, avec tous les résultats étiquetés correctement. Cependant, aucun modèle ne se distingue comme un bon prédicteur de notre ensemble de données.

Pour entrer dans plus de détails sur les performances du modèle, nous avons choisi l'un des modèles les plus performants : la forêt aléatoire qui génère une forêt d'arbres de décision de taille et de profondeur variables. Le paramètre optimal (calculé par un algorithme *ad*-hoc de recherche de paramètres optimaux) pour la forêt aléatoire est de 140 arbres avec une profondeur maximale de deux, ce qui correspond à un paramètre moyen pour une telle propriété de jeu de données (en termes de nombre de données et de structure de données). La précision des détails du modèle de forêt aléatoire sur la figure 3 montre deux phénomènes distincts. D'une part, le modèle peut prédire quelque chose lorsque les sites Web appartiennent au cluster 2 ou au cluster 1 (les deux clusters les plus peuplés), même si sa précision à le faire est très faible. En fait, le rappel de classe de la classe 1 est de 80% tandis que le rappel de classe de la classe 2 est de près de 70%, ce qui indique que le modèle renvoie des prédictions réelles pour ces classes (il se trompe peu). Cependant, leur

précision de faible classe indique que même si nous pouvons prédire quelque chose pour eux, le modèle n'est pas très sélectif et faire une vraie prédiction est toujours difficile à faire avec une moyenne de précision de 42% (35,94 + 48,33) / 2. Pour résumer, de nombreux résultats sont renvoyés mais la plupart des prédictions sont incorrectes. En revanche, ce n'est pas vrai du tout pour les 3 autres classes. En fait, aucun des 3 autres clusters (classe 3,4,5) n'a été prédit par le modèle de forêt aléatoire. Cela indique que les balises que nous avons créées ne fonctionnaient que sur deux clusters mais ne sont pas suffisamment spécifiques pour vraiment prédire ces deux clusters alors que les 3 clusters restants ne sont tout simplement pas représentés par les balises. Ces derniers sont en effet trop hétérogènes dans la distribution de leurs balises et pas assez spécifiques pour qu'un groupe de balises soit prédit avec précision.

accuracy: 41.93% +/- 4.00% (micro average: 41.94%)

	true class2	true class4	true class1	true class3	true class5	class precision
pred. class2	23	7	7	20	7	35.94%
pred. class4	0	0	0	0	0	0.00%
pred. class1	10	9	29	3	9	48.33%
pred. class3	0	0	0	0	0	0.00%
pred. class5	0	0	0	0	0	0.00%
class recall	69.70%	0.00%	80.56%	0.00%	0.00%	

Figure 3 Performances du modèle de forêt aléatoire pour prédire chaque cluster en fonction des balises.

Le résultat montre que le corpus en ligne du diabète a son propre espace et est organisée en communautés. Il est remarquable que ces communautés sont révélées par des caractéristiques liées aux nœuds et aux arrêtes d'un graphe tandis que ces communautés partagent, dans la vie réelle, un intérêt commun que l'on devine en examinant leur contenu si l'on a pas pu le prouver formellement en l'état actuel de nos travaux.

Cependant, on constate une faible performance de prédiction en utilisant des balises pour fournir une explication sémantique des grappes de sites Web liés au diabète obtenues dans nos travaux précédents. En regardant la distribution des balises, ce résultat reflète le fait que certaines

balises sont spécifiques au cluster 1 ou 2 et ce sont celles avec le poids maximum trouvé à partir de 7 méthodes. Cela signifie que deux de nos clusters ont une identité relativement facile à prévoir avec des balises. Cependant, même pour le cluster 1 ou 2, il est difficile d'expliquer la prédiction par une combinaison spécifique des balises (une signature sémantique).

De tels résultats montre une faible performance prédictive en utilisant des balises pour fournir une explication sémantique des clusters de Dia-Map. Il découvre la réalité d'une communauté Web : un mélange de sites Web de différents types qui créent un espace mixte mais localisé. Cela prouve également que la communauté peut avoir un schéma de marquage de temps en temps, mais il est toujours difficile d'utiliser une approche sémantique pour prédire avec précision les clusters. Néanmoins, l'approche sémantique est appropriée pour la recherche d'informations de qualité à condition de fortement les raffiner.

Applications réelles

Nous avons choisi Google, Bing, Baidu et Yahoo, les quatre moteurs de recherche généralistes les plus populaires pour faire une comparaison entre les réponses des moteurs et ceux de *DiaMap* sur un échantillon de questions sur le diabète. Le principe est de choisir 5 questions sur le diabète potentiellement intéressantes pour un utilisateur réel et de chercher des informations intéressantes sur des sites correspondant sur les moteurs de recherches et de manière équivalente, en traduisant les balises correspondantes en tags sur *DiaMap*. La comparaison se fera entre les sites proposés par les moteurs de recherche et les sites proposés par *DiaMap* selon deux caractéristiques :

• La qualité des résultats de recherche estimés selon un expert en diabète traduit en pertience;

• la quantité de site pertinents rapportée par les moteurs de recherche, par *DiaMap*, et combien de ces sites sont partagés entre les deux corpus résultats.

Les 5 questions[1] ont été proposées par un expert en diabète qui a également collecté ces données à partir de cas d'utilisateurs réels. Un comité composé de l'expert et deux chercheurs a examiné et décidé des questions finales qui couvrent les blogs, les sujets de niche du diabète comme la dia-boulimie, les achats en ligne, les informations hospitalières et enfin les organisations caritatives. Ces 5 thèmes et leurs questions associées sont soumis aux moteurs de recherche desquels seuls les 5 premiers résultats sont retenus; ce, en vertu de la très faible propension des utilisateurs des moteurs à aller au-delà du 5ème résultat, que seuls 10% dépasseront. L'expert en diabète lit ensuite ces résultats et décide si les résultats sont liés ou non aux thématiques demandés. Il en résulte un corpus « RRSE » de résultats pertinents du moteur de recherche et un indicateur « RRSECount » représentant le nombre de résultats pertinents pour la question par un moteur de recherche donné.

Dans un second temps, nous procédons à une opération similaire avec *DiaMap* pour extraire les sites dont les tags correspondent aux questions et on obtient un corpus de sites « WDiamap », ainsi qu'un deuxième indicateur « WDiamapCount », soit le nombre de sites Web appropriés à chaque question, présent dans *DiaMap*, et répondant au tag correspondant. Ces sites sont pertinents car déjà sélectionnés par un expert suivant une procédure décrite dans la méthodologie de création de *DiaMap*. Le troisième indicateur « RRSEintersectWDiamap » relève combien de résultats pertinents sont partagés entre les moteurs de recherche et *DiaMap*. Pour finir nous avons ajouté deux indicateurs contextuels : « SubMapCount » est le nombre de sites internet de *DiaMap* liés à ceux de la WDiamap ; et « RRSEinSubMap » cherche dans le sous espace de *DiaMap* ainsi isolé si les sites de RRSE s'y trouvent.

Tout ceci a pour objectif de vérifier que *DiaMap* peut offrir aux utilisateurs des informations de qualité, contextualisées dans un sous-espace du diabète correspondant tandis que les moteurs de recherche donnent un accès ponctuel à un information décontextualisée.

[1] Les 5 questions sont : blogs pour le support psychologique en diabète gestationnel ; existe-t-il des associations traitant de la diaboulimie ; où trouver un magasin en ligne pour la décoration de pompe à insuline ; y-a-t-il des hopitaux ou cliniques dédiés au diabète type 1 ; Je veux faire un don à des organisations sur le diabète, quels sont les sites de dons.

Les sous-cartes visualisées montrent les sites Web sources initiaux de *DiaMap* correspondant à chaque scénario ainsi que les sites Web liés entre eux. Nous pouvons vérifier dans cette carte si tout ou partie des sites Web proposés par les moteurs de recherche figurent dans les sous-cartes ou non.

Avec les 5 scénarios ou 5 thématiques, *DiaMap* a obtenu une moyenne de résultat par question toujours supérieur aux moteurs de recherches exception faite de la question de la vente en ligne où *DiaMap* est en deçà de la moyenne des moteurs de 0.5 point. Seuls deux sites sur les résultats des 5 questions ont été proposé en commun. Les résultats des sous-cartes sont relativement fournis avec plus de 40 sites à chaque fois à l'exception des de la vente en ligne, point fort des moteurs. En guise de résultat final, les sites Web principaux et précis de *DiaMap* peuvent répondre à des questions précises et privilégient des utilisateurs souhaitant approfondir le sujet. Seuls deux sites proposés par les moteurs n'étaient pas au moins présents dans notre corpus et donc répertoriés.

Les moteurs de recherche quant à eux peuvent offrir un éventail de sites Web aux utilisateurs pour répondre à leurs questions. Cependant, il faut que les utilisateurs lisent chaque site Web et doivent encore juger si les sites Web sont fiables alors qu'ils ne disposent souvent pas d'informations pour cette évaluation. *DiaMap* ressemble plus à un contenu organisé par des communautés en ligne sur le diabète.

Avec cet outil, nous pouvons aider les gens à trouver des sites Web plus précis et centrés sur leur thématiques, en utilisant des balises et des classes pour les distinguer.

En résumé, *DiaMap* présente la visualisation d'informations en forme de carte de sites Web liés au diabète pour manifester le contenu en ligne lié au diabète et sa structure. Différent des moteurs de recherche traditionnels, *DiaMap* présente l'ensemble des communautés et utilise des balises pour identifier les sites Web plus pertinents et plus fiables. Il modifie en quelque sorte la façon de naviguer dans les informations en ligne sur le diabète et pourrait être une alternative pour augmenter les moteurs de recherche actuels. *DiaMap* peut devenir un moteur de recherche qui se

concentre sur le monde en ligne mondial du diabète pour aider les personnes atteintes de leur diabète.

À l'avenir, nous augmenterons la taille de l'ensemble de données jusqu'à plus de 5000 sites Web dans *DiaMap*. Notre idée est d'utiliser le traitement automatique du langage naturel (TALN) pour apprendre les balises les plus pertinentes des sites Web et annoter automatiquement un corpus beaucoup plus vaste. Cette étude conclut que l'espace en ligne sur le diabète existe et que ce nouveau monde virtuel en tant que ressource ouverte devrait être connu de tous dans son entièreté. Ceci est juste le point d'entrée pour présenter comment combiner la visualisation de réseau et la détection communautaire à la fois topologique et sémantique pour faire de l'analyse de réseau pour le diabète. Cette nouvelle approche peut également être étendue à l'analyse de réseaux pour les autres maladies chroniques aidant les patients et leurs familles à s'orienter dans un monde numérique traditionnellement opaque.

Contents

Acronyms

- **American Diabetes Association (ADA)**
- **Alternative for Germany" (AFG)**
- **Continuous Glucose Monitoring (CGM)**
- **Diabetes Attitude, Wishes and Needs Study (DAWN)**
- **Diabetes ketoacidosis (DKA)**
- **Diabetes self-management education (DSME)**
- **Diabetes Mellitus (DM)**
- **Gestational diabetes mellitus (GDM)**
- **Health Terminology/Ontology Portal (HeTOP)**
- **Hyperlink Network Analysis (HNA)**
- **Hyper Text Markup Language (HTML)**
- **International Diabetes Federation (IDF)**
- **Impaired Glucose Tolerance (IGT)**
- **Medical nutrition therapy (MNT)**
- **Non-government organizations (NGOs)**
- **National Health Service (NHS)**
- **Nature Language Processing (NLP)**
- **Principal Component Analysis (PCA)**
- **Relevant Results Search Engine (RRSE)**
- **Relevant Results Search Engine Count (RRSEcount)**
- **Relevant Results Search Engine Intersect Website DiaMap (RRSEintersectWDiaMap)**
- **Relevant Results Search Engine in {W-DiaMap} hyperlinks submap (RRSEinSubMap)**
- **Search Engine (SE)**
- **Self-monitoring of blood glucose (SMBG)**
- **{W-DiaMap} hyperlinks submap count (SubMapCount)**
- **Type 1 diabetes mellitus (T1DM)**
- **Type 2 diabetes mellitus (T2DM)**
- **World Diabetes Foundation (WDF)**
- **World Health Organization (WHO)**
- **University College London Hospitals (UCLH)**
- **Websites DiaMap Count (WDiaMapCount)**
- **World Wide Web (WWW)**

1 Chapter 1 Introduction

Diabetes is one of the largest global health emergencies of the 21st century. It is a chronic condition which requires continuous medical care and constant patient self-management (Shaw, Sicree, & Zimmet, 2010). As a patient having diabetes for more than 15 years, I used to work for International Diabetes Federation (IDF) being an ambassador to do advocacy on improving social awareness in China. During the volunteer work, I gained a lot of first-hand information about people living with diabetes. In 2015, I was invited to deliver a speech talking about the current situation of people with diabetes in China. I noticed, in addition to discussing the treatments options for diabetes, that there are still lots of social facts influencing the quality of patients 'life. I got inspired by one study initiated by Novo Nordisk[2] named DAWN study. DAWN study is about the Diabetes Attitudes, Wishes, and Needs Study. Although data on the benefits of near-normal glucose control is widely accepted and the treatments for diabetes care are more efficient and available than before, the results are still not optimal. In fact, factors other than knowledge and effective therapy affect the behavior of patients and health professionals and influence their ability to make optimal use of existing treatments. The purpose of the DAWN study is to determine the broad attitudes, aspirations and needs of people with diabetes and caregivers, and to lay the foundation for efforts to improve diabetes care. The study evaluated several factors associated with the quality of diabetes care: diabetes self-management levels and psychological distress, the quality of relationships between people and diabetes care providers, collaboration between diabetes care providers, and effective drug therapy (Skovlund & Peyrot, 2005).

Such care involves several stakeholders as hospitals, physicians, research laboratories, pharmaceutical, companies, medical and social relay structures, associations, health-related NGOs Non-Governmental Organizations (NGOs), patients and their families, publications around diabetes care. All of these key diabetes actors play their own roles which create the ecosphere of diabetes. How to figure the relationship among them? How to display the network of stakeholders in diabetes world? With these kinds of questions in my mind, I met my co-supervisor Fabien Pfaender who is working in the "Connaissance Organisation et Systèmes TECHniques (Costech)"

[2] Novo Nordisk is a global healthcare company with more than 95 years of innovation and leadership in diabetes care, headquartered in Denmark.

laboratory by UTC university. The UTC-Costech Laboratory is a pluridisciplinary research unit where research scientists work on the triple interface "Man, Technology, Society" for the purpose of describing scientifically, analyzing, modeling and designing tooled interactions in complex social, technical situations. The area explored constitutes a crossroad for engineering sciences, philosophy and social sciences (cognition, economics & management). The Costech laboratory analyses the technical aspects of these interactions in human experience and social practice. Fabien runs a special program in Shanghai officially called "Cyber-physical systems and data science lab". After we met in Shanghai, he introduced Marie-Christine Jaulent as my director. Marie-Christine is in charge of the "laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé (LMICS)" in Paris. LIMICS is a single-team, interdisciplinary research unit in computer science and medical informatics. My master degree was major in Communications and mainly focused on what are the International non-government organizations (NGOs) running models. Being neither a computer scientist nor a clinic physician, my approach was to use the state-of-art tools from computer science discipline to help people find a better way to access information of diabetes online and also to show relationship of diabetes-related websites in the digital world.

1.1 About diabetes in general

Diabetes mellitus (DM) describes a group of metabolic disorders characterized by increased blood glucose concentration. It is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. There are three main types of diabetes as type 1, type 2 and gestational diabetes ("About Diabetes", World Health Organization, 2014).

Common symptoms of diabetes include: urinating often, feeling thirsty, feeling hungry even though you are eating, extreme fatigue, blurry vision, cuts/bruises that are slow to heal. Symptoms of type 1 diabetes often appear suddenly while some people with type 2 diabetes have symptoms so mild that they go unnoticed. Although there are similarities between type 1 and type 2 diabetes, their causes as well as their treatment are different (Roglic & World Health Organization, 2016). Gestational Diabetes is high blood sugar that develops during pregnancy and usually disappears after giving birth. It can cause problems for women and babies during

pregnancy and after birth. But the risks can be reduced if the condition is detected early and well managed.

1.1.1 Type 1 diabetes mellitus (T1DM)

T1DM, also known as juvenile diabetes or insulin-dependent diabetes is usually diagnosed in children, teens, and young adults which the pancreas produces very little or no insulin. It is an auto-immune system condition in which the immune system is activated to destroy the cells in the pancreas which produces insulin (Atkinson, Eisenbarth, & Michels, 2014).

Globally, the incidence and prevalence of T1DM vary substantially. The majority of incidence comes from Europe and North America. T1DM is most common in Finland (>60 cases per 100 000 people each year) and Sardinia (around 40 cases per 100 000 people each year). According to the resources from website "Beyond Type1" (https://beyondtype1.org/type-1-diabetes-statistics/, access by September 2019), we can see below the statistics of T1DM in France and China.

France

- Over 7% of France's adult population of 45 million have diabetes.
- In 2013, France was listed as #21 among countries according to rate of incidence of Type 1 diabetes among children 0-14 years old, with a rate of 12.2 out of 100,000.
- The rate of incidence of Type 1 increased to about 18 per 100.000 people as of 2015.
- The highest rates of incidence occurred in the regions of Corsica (21.7 per 100,000 people), Provence-Alpes-Côte d'Azur (21.1 per 100,000) and Hauts-de-France (19.7 per 100,000).
- France does not yet have a national register for Type 1 diabetes.

China

- Nearly 50,000 children and adolescents under age 20 are living with Type 1diabetes in China, placing the country in the top 10 countries according to the eighth edition of the IDF Diabetes Atlas (Ogurtsova et al., 2017).
- Dr. Weng Jianping led a study that found that China is among the countries with the lowest rates of incidence of Type 1 diabetes in children and adults.

- Over 60% of new diagnoses over a three-year period were among people 20 years old or older.
- Dr. Weng's study concluded that populations in Northern China experience a higher rate of incidence of Type 1 diabetes than regions in the south.
- The same study also found that in 2015, China had the largest estimated number of new annual cases of Type 1 diabetes in children of any country in the Western Pacific, at a rate of over 4,000 out of 10,000.
- China does not yet have a national register for Type 1 diabetes.

The mechanisms underlying these figures of geographical incidence of Type 1 diabetes are unknown, but have largely been attributed to environmental influences and genes (Todd, 2010).

1.1.2 Type 2 diabetes mellitus (T2DM)

T2DM is the most common form of diabetes, accounting for around 90% of all diabetes. It is often linked to being overweight or inactive, or having family history of type 2 diabetes (Olokoba, Obateru, & Olokoba, 2012). In T2DM, body does not use insulin properly. This is called insulin resistance. At first, the pancreas makes extra insulin to make up for it. But, over time the pancreas isn't able to keep up and can't make enough insulin to keep blood glucose levels normal. Type 2 diabetes is most commonly diagnosed in older adults, but is increasingly seen in children, teenagers and younger adults due to rising levels of obesity, physical inactivity and poor diet (Association, 2000, p. 2).

T2DM was relatively rare in developing countries some decades ago; for example, the prevalence of the disease was <1% in China in 1980. However, the major burden of diabetes mellitus is now taking place in developing rather than in developed countries. 80% of cases of diabetes mellitus worldwide live in less developed countries and areas. Asia has emerged as the "diabetes epicenter" in the world, as a result of rapid economic development, urbanization and nutrition transition over a relatively short period of time (Chan et al., 2009).

Among the 10 countries with the largest numbers of people predicted to have diabetes mellitus in 2030, five are in Asia (China, India, Pakistan, Indonesia and Bangladesh). (see figure

1) In particular, the latest figures derived from a national survey in China between 2007 and 2008 suggest that China has overtaken India and become the global epicenter of the diabetes epidemic with more than 92 million adults (9.7% of the total population) with diabetes mellitus (Yang et al., 2010).

2011		2030	
Country	**Millions**	**Country**	**Millions**
China	90.0	China	129.7
India	61.3	India	101.2
United States of America	23.7	United States of America	29.6
Russian Federation	12.6	Brazil	19.6
Brazil	12.4	Bangladesh	16.8
Japan	10.7	Mexico	16.4
Mexico	10.3	Russian Federation	14.1
Bangladesh	8.4	Egypt	12.4
Egypt	7.3	Indonesia	11.8
Indonesia	7.3	Pakistan	11.4

Figure 1 TOP 10 countries for numbers of people aged 20-79 years with diabetes in 2011 and 2030 (Whiting, Guariguata, Weil, & Shaw, 2011).

Compared with developed countries, the proportion of young to middle-aged individuals with T2DM is higher in developing countries. Furthermore, it is not necessarily less prevalent in rural than in urban areas of developing countries, as is generally believed. The rural-urban difference in prevalence is predicted to narrow owing to urbanization, rural to urban migration and its associated lifestyle changes. A study from India showed a significant increase in diabetes mellitus prevalence in both urban (from 13.9% in 2000 to 18.2% in 2006) and rural areas (from 6.4% in 2000 to 9.2% in 2006) (Ramachandran, Mary, Yamuna, Murugesan, & Snehalatha, 2008). Similar findings have been reported from other Asian countries.

Although obesity remains a key driver of T2DM, several other factors are attributable to the diabetes epidemic other than obesity. They include fetal and early life nutrition status, as well

as some factors associated with rapid socioeconomic development, such as depression, sleeping disorders and environmental pollutants (Alonso-Magdalena, Quesada, & Nadal, 2011).

1.1.3 Gestational diabetes mellitus (GDM)

GDM is the third main form of diabetes which occurs when pregnant women develop high blood sugar levels without a previous history of diabetes. It is defined as any degree of glucose intolerance with onset or first recognition during pregnancy (Metzger, 1998). This definition applies whether insulin or only diet modification is used for treatment and whether the condition persists after pregnancy. It does not exclude the possibility that unrecognized glucose intolerance may have antedated or begun concomitantly with the pregnancy.

Approximately 7% of all pregnancies are complicated by GDM, resulting in more than 200,000 cases annually. The prevalence may range from 1 to 14% of all pregnancies, depending on the population studied and the diagnostic tests employed (Association, 2004).

1.1.4 Global prevalence of diabetes

The global prevalence of diabetes and impaired glucose tolerance in adults has been increasing over recent decades.

In 1980, the World Health Organization (WHO) estimated that there were 108 million people living with diabetes and this number increased fourfold in 2014 (Roglic & World Health Organization, 2016). The International Diabetes Federation (IDF) estimated the global prevalence to be 194 million in 2003, 246 million in 2006, 285 million in 2009, 366 million in 2011(Whiting et al., 2011), 382 million in 2013 (Guariguata et al., 2014), and 415 million in 2015 (Ogurtsova et al., 2017). Figure 2 shows the number of 7 regions from IDF, North America and Caribbean, Europe, Middle East and North Africa, Western Pacific, South East Asia, Africa, South and Central America, people with diabetes in 2015 and predicted number in 2040.

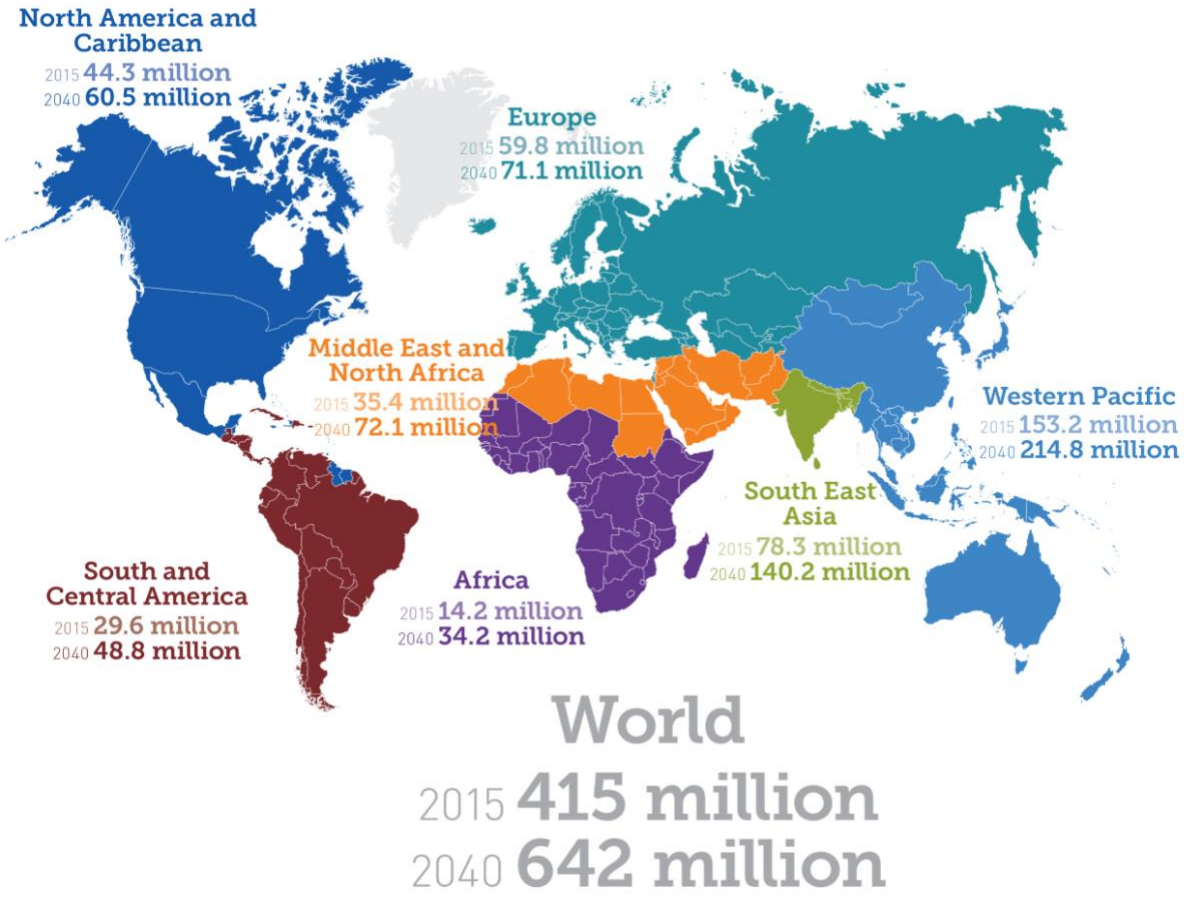


Figure 2 Estimated number of people with diabetes worldwide and per region in 2015 and 2040 (20-79 years) (Ogurtsova et al., 2017).

According to the latest report from IDF, it was estimated that in 2017 there are 451 million (age 18-99years) people with diabetes worldwide. These figures were expected by the same source to increase to 693 million by 2045. It was estimated that almost half of all people (49.7%) living with diabetes are undiagnosed. Moreover, there was an estimated 374 million people with impaired glucose tolerance (IGT) and it was projected that almost 21.3 million live births to women were affected by some form of hyperglycemia in pregnancy (Cho et al., 2018).

1.2 Current treatment and management of diabetes

People living with diabetes have a higher risk of morbidity and mortality than the general population. Diabetes is one of the leading causes of premature death worldwide (Susan van,

Beulens, Yvonne T. van Der, Grobbee, & Nealb, 2010). According to a 2016 World Health Organization report, an estimated 1.6 million deaths were directly caused by diabetes. People with diabetes have increased risk of developing several serious life-threatening health problems resulting in higher medical care costs, and reduced quality of life. Persistent high blood glucose levels cause generalized affecting the heart, eyes, kidneys and nerves and resulting in various complications (Organization, 2016). Hence, diabetes is also the major cause of blindness, renal failure, heart attacks, stroke and lower limb amputations. The global healthcare expenditure on people with diabetes was estimated to be USD 850 billion in 2017 (Ogurtsova et al., 2017).

As there is still no cure for diabetes so far, we only can manage diabetes with current knowledge. However, effective approaches are available to prevent type 2 diabetes and to prevent the complications and premature death that can result from all types of diabetes.

1.2.1 Management of T1DM

There is no way to prevent the occurrence of T1DM so far. However, the discovery of insulin in 1921–1922 was clearly the most significant therapeutic event in the history of T1DM. Before the discovery of insulin in 1921, there was no better way to reduce the blood sugar of patients with T1DM. Most of the patients died of various complications of diabetes. With the discovery and application of insulin, patients with T1DM can enjoy life just like normal people. Usually, Insulin therapy is given by injection under the skin. But in modern countries, it combines with mechanical technologies (e.g., insulin pumps) for improved treatment (Blumer, Edelman, & Hirsch, 2012).

In addition to improved insulin preparations and delivery systems, self-monitoring of blood glucose (SMBG) is also playing an important role in managing type 1 diabetes. Usually people with T1D are advised to check 7 times their blood sugar per day. The new technology real-time continuous glucose monitoring (CGM) can help patients monitor their blood sugar more often and easily (Hirsch, 2009).

With insulin pumps and continuous glucose monitoring improving diabetes care, these two technologies are now being used together as sensor-augmented pump therapy (Bergenstal et al.,

2011). Although current sensor-augmented pump therapy uses each device independently, integration of both systems is being investigated. A key element for such efforts involves low-glucose suspend systems that monitor blood glucose with a continuous glucose monitor and suspend insulin delivery when glucose falls below a preset threshold for up to 2h, to prevent hypoglycemic episodes (Hirsch, 2012). In the future, artificial pancreas which combine the implantable insulin pump and continuing glucose monitor will ease the lifestyles in addition to preventing complications.

Self-management of type 1 diabetes behaviors include collaboration with parents, diabetes care activities, diabetes problem solving, diabetes communication, and goals. The goals of self-management are to optimize the blood sugar control, prevent acute and chronic complications, and optimize quality of life, while keeping costs acceptable. Since type 1 diabetes usually happens before 18 years old, how to educate parents dealing with children with diabetes is also very crucial. From the insulin treatment, blood glucose testing, food and beverage offering, moderating exercises to psychology support, type 1 diabetes needs more attention from daily-based care. For example, kids need efficient calories to grow up, but meanwhile, they still need to prevent from taking too much calories for better blood sugar level. They probably do more physical activities than people with type 2 diabetes, at the same time, hypoglycemia needs to be prevented. How to manage type 1 diabetes is related to take balance with diet and excises on a daily-base, in that case, joining in type 1 communities and sharing blogs are especially important during the Diabetes self-management education (DSME).

1.2.2 Management of T2DM

T2DM is treated with lifestyle changes, oral medications (pills), and insulin. Some people with type 2 can control their blood glucose with healthy eating and being active. But, oral medications or insulin may be needed to meet the target blood glucose levels (Sibal & Home, 2009, p. 2). Type 2 usually gets worse over time – even when the patients don't need medication at first, they may need it later on.

Different from T1DM, T2DM is preventable. Individuals with blood glucose levels higher than normal but not high enough for a diagnosis of T2DM, such as those with Impaired Glucose

Tolerance (IGT), are usually considered to have a high risk of future T2DM. Several major trials show that intensive lifestyle interventions, specifically aimed at weight loss and increased physical activity in high-risk individuals, significantly reduced conversion to diabetes in high-risk patients with IGT by 58% (Knowler et al., 2002). Prevention of T2DM is a lifelong task and requires an integrated approach combined with five major lifestyle factors: diet, physical activity, smoking, overweight or obesity (Hu, van Dam, & Liu, 2001).

Metformin (as usual first-line therapy) or sulfonylurea are usually recommended by global guidelines for T2DM (International Diabetes Federation, 2005). Self-monitoring of blood glucose (SMBG) is also part of caring. As it is less demanding as T1DM, SMBG was associated with decreased diabetes-related morbidity and all-cause mortality in T2DM, and this association remained in a subgroup of patients who were not receiving insulin therapy (Martin et al., 2006).

As the evidence suggests that lifestyle management can have profound effects on control of blood glucose, nutritional advice is playing an important part in managing T2DM. Dietetic consultation should be provided to all newly diagnosed people and reviewed annually, using the skills of someone with specific expertise in the field. The advice itself in general follows that of healthy balanced eating with encouragement of high-fiber, low glycemic index sources of carbohydrate and oily fish, and control of the intake of foods containing free sugars and saturated and trans fatty acids (Hu et al., 2001). For patients with diabetes, a meal planning system providing consistency in the carbohydrate content of meals is recommended. Since there are various kinds of food and countless combinations, people with T2DM better personalized their meals according to their own diet habits. Effective diabetes education is definitely an integral part of comprehensive diabetes care.

1.2.3 Management of GDM

All women with GDM should receive nutritional counseling, by a registered dietitian when it is possible. Individualization of medical nutrition therapy (MNT) depending on maternal weight and height is recommended. MNT should include the provision of adequate calories and nutrition to meet the needs of pregnancy and should be consistent with the maternal blood glucose goals that have been established. No caloric sweeteners may be used in moderation (Major, Henry, de Veciana, & Morgan, 1998).

Insulin is the most prescribed drug therapy to reduce the incidence of fetal disease. When maternal glucose levels are used, insulin therapy is recommended when MNT fails to maintain self-monitored glucose at the following levels ("A Comparison of Glyburide and Insulin in Women with Gestational Diabetes Mellitus | NEJM," n.d.).

Programs of moderate physical exercise have been shown to lower maternal glucose concentrations in women with GDM. Although the impact of exercise on neonatal complications awaits rigorous clinical trials, the beneficial glucose-lowering effects warrant a recommendation that women without medical or obstetrical contraindications be encouraged to start or continue a program of moderate exercise as a part of treatment for GDM (Brankston, Mitchell, Ryan, & Okun, 2004). Breast-feeding, as always, should be encouraged in women with GDM (Mayer-Davis et al., 2006)(Schaefer-Graf et al., 2006).

There are many secondary preventions as glucose control and tertiary prevention like regularly checking for eye, foot and kidney abnormalities can improve the outcomes for people with diabetes.

However, the self-management is still playing the main role in avoiding or delaying the complications. Training and educating as well as accessing up to date knowledge to help people self-manage their diabetes helps prevent unnecessary health care utilization and hospitalization (Kent et al., 2013).

Managing diabetes as a life-long condition concentrates on keeping blood sugar levels as close to normal, without causing low blood sugar. Studies have shown that people with diabetes have a lower incidence of complications if they maintain good blood glucose performance (Poolsup, Suksomboon, & Rattanasookchit, 2009). That's usually associated with a complex set of services and support ranging from glucose monitoring, insulin and other medication management (insulin in the case of type 1 diabetes; oral medications, as well as possibly insulin, in type 2 diabetes), psychotherapy and social support, to physical activity promotion, nutrition counselling and more. Integrating these supports into a patient's therapeutic regimen presents challenges that need to be addressed through a variety of strategies. People with diabetes can benefit from education about the disease and treatment, good nutrition and exercise to achieve the goal of keeping both short-term and long-term blood glucose levels within the ideal range. In

addition, given the associated higher risks of cardiovascular disease, lifestyle modifications are recommended to control blood pressure (Haw et al., 2017).

1.3 The growing role of Internet to seek for information in self-management

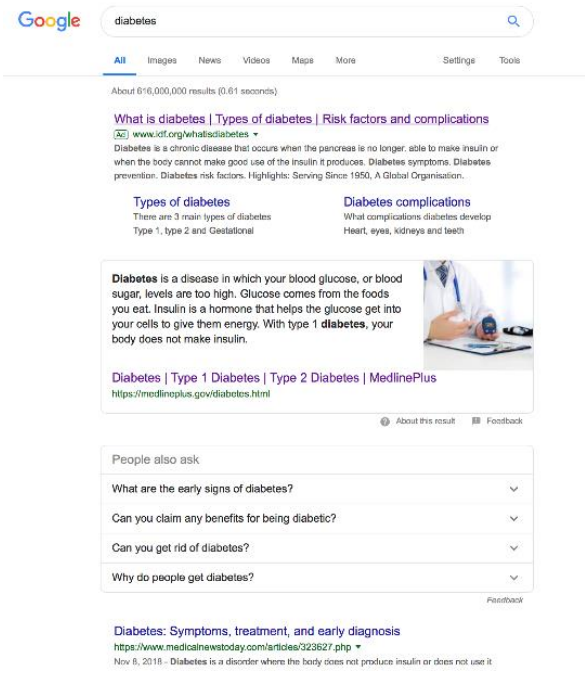
With the rapid development of the Internet, the WWW has become a carrier of a large amount of information leading Internet to become a favored source to find health information. Worldwide, about 4.5% of all Internet searches are for health-related information (Eysenbach & Kohler, 2003). Searching for health information is now the third most popular use of Internet technology (Fox & Fallows, 2003). People having a chronic disease like diabetes, are more likely to engage intensely with online resources (Fox, n.d.). Research demonstrated that online information targeting chronic disease has been shown to improve health outcomes and decrease the utilization of health care resources (Kate R. Lorig, Philip L. Ritter, Ayesha Dost, Kathryn Plant, Diana D. Laurent, Ian Mcneil, 2008). Nowadays, more and more people go through the internet to find online resources to get health information and gaining social support. Most people, looking for online health information, typically use general search engines to find specific health conditions and enter short sentences, often misspelled. They seldom go beyond the first 2 pages of a search result. Both their search and evaluation skills are limited although they are concerned about the quality of online health information (Morahan-Martin, 2004).

In order to illustrate what is previously said, we used 2-popular search engines, Google and Yahoo, with typing the keyword “diabetes”. The result of the search shows the TOP 10 websites on the first page of Google (see figure 3):

1. https://www.idf.org
2. https://www.medlineplus.gov
3. https://www.medicalnewstoday.com
4. http://www.diabetes.org
5. https://www.webmd.com
6. https://www.medicinenet.com
7. https://en.wikipedia.org
8. https://www.nhs.uk

9. https://www.mayoclinic.org
10. https://www.nhsinform.scot

www.idf.org is an umbrella organization of over 240 national diabetes associations in 168 countries and territories. www.medineplus.gov is the United States National Institutes of Health's Web site for patients and their families and friends. It is produced by the world's largest medical library, bringing the information about diseases, conditions, and wellness issues. www.medicalnewstoday.com is an UK Health line media, which is for publishing the general medical information. www.diabetes.org is the American Diabetes Association. www.webmd.com is a platform offering the quality, accuracy and security of general medical information. www.medicinenet.com is a website offering the general health solutions. en.wikipedia.org is the free encyclopedia in English. www.nhs.uk is the UK's biggest health website with more than 43 million visits per month. www.mayoclinic.org is a clinic belonging to Mayo Foundation for Medical Education and Research. www.nhsinform.scot is Scotland's national health information service.



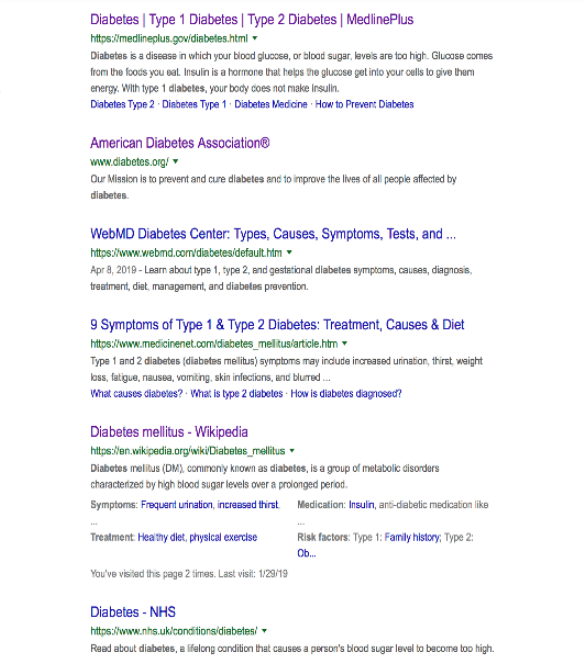


Figure 3 The TOP 10 websites showing on the first page of Google by typing keyword "diabetes" (Accessed until February 2020).

It's not difficult to find that the top 10 websites returned by Google are more world-leading general medical information websites instead of some specific topic domain websites like diet, exercises or psychology support, etc.

By using Yahoo, the TOP 11 websites showing on the first page are (see figure 4):

1, https://www.lifeextension.com
2, https://www.sellyourstripsformore.com
3, https://www.diabetes.org
4, https://www.webmd.com
5, https://www.medlineplus.gov
6, https://en.wikipedia.org
7, https://www.medicalnewstoday.com
8, https://www.nhs.uk
9, https://www.mayoclinic.org
10, https://www.medicinenet.com
11, https://www.cdc.gov

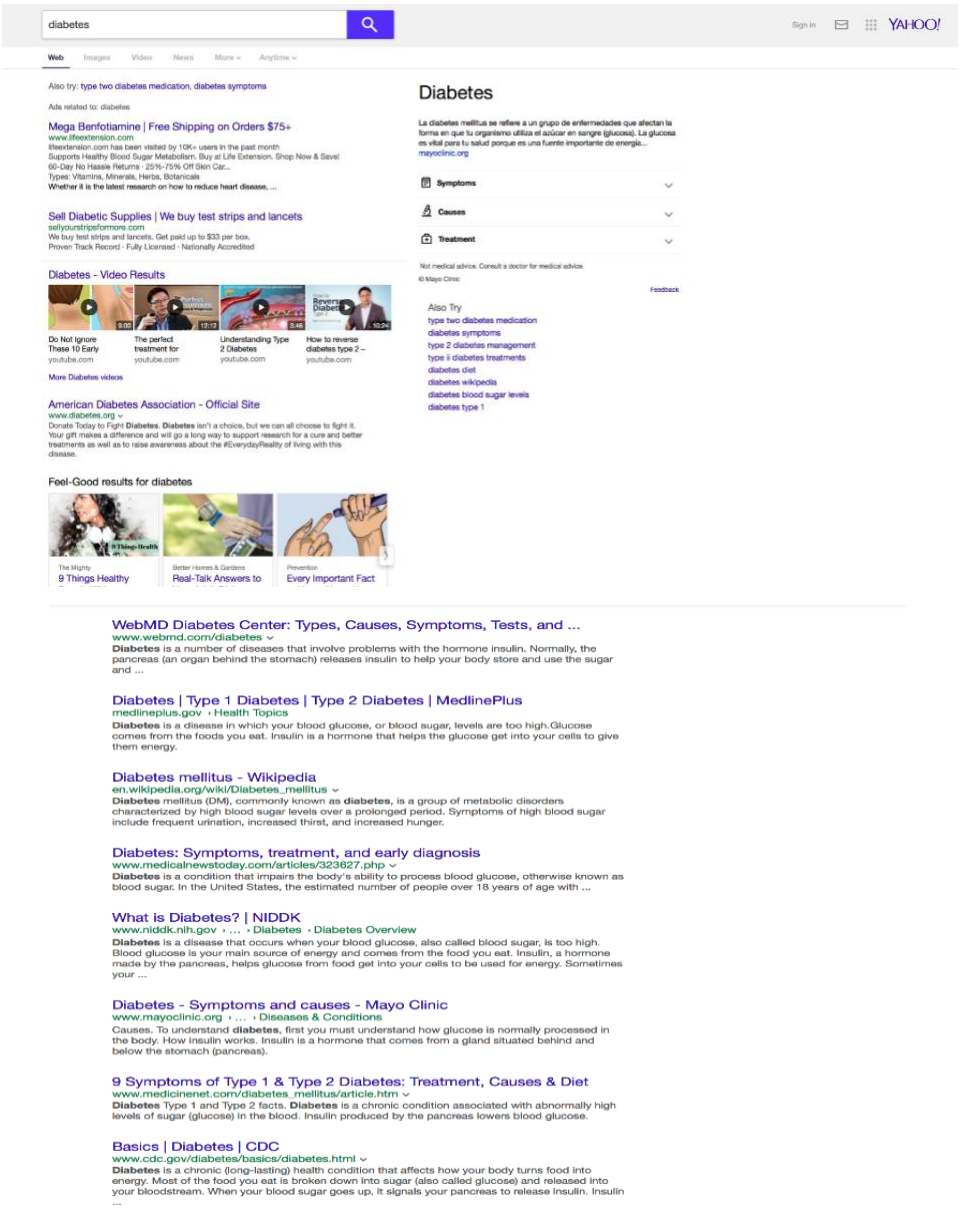


Figure 4 The TOP 11 websites showing on the first page of Yahoo by typing keyword "diabetes" (Accessed until February 2020).

Surprisingly, the top 2 websites recommended by Yahoo are online shops. www.lifeextension.com mainly sells the nutritional supplements and www.sellyourstripsformore.com is to sell diabetes test strips for cash. The other 8 websites www.diabetes.org, www.webmd.com, www.medlineplus.gov, en.wikipedia.org, www.medicalnewstoday.com, www.nhs.uk, www.mayoclinic.org, www.medicinenet.com are exactly the same as the results from Google search engine with a different order. The last one

showing on the first page, www.cdc.gov, is American center for disease control and prevention. Like with Google search engine, the websites related keyword “diabetes” are more world-leading general medical information domain instead of specific forums, blogs or online communities, etc.

Both results are shown by typing keyword “diabetes” in the search engines. Due to the algorithms behind them such as Google and Yahoo, the results depend on the individual’s search and evaluation skills. It also strongly depends on the search criteria entered. Currently, PageRank is a commonly used algorithm to rank web pages in their search engine results. PageRank is a way of measuring the importance of website pages. According to Google, PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the websites are likely to receive more links from other websites (Wills, 2006).

Here is another example with search engines. If people just diagnosed type 1 diabetes want to find any hospital or clinic dedicated to T1DM, we assume he/she typed “Is there any hospital or clinic dedicated to type 1 diabetes” in Google. The top 10 websites showing on the first page of Google are below:

1, https://www.armi.org.au
2, https://www.idf.org
3, https://www.em-consulte.com/en/article/276237
4, https://www.lewishamandgreenwich.nhs.uk/diabetes-services-in-lewisham/
5, https://www.uclh.nhs.uk
6, https://www.worlddiabetesfoundation.org
7, https://www.ncbi.nlm.nih.gov/books/NBK343389/
8, https://onlinelibrary.wiley.com/doi/abs/10.1111/imj.13649
9, https://www.joslin.org
10, https://www.royalfree.nhs.uk

www.armi.org.au is the Australian Regenerative Medicine Institute opened in 2009, mainly contributions to regenerative medicine and stem cell research. www.idf.org is an umbrella organization of over 240 national diabetes associations in 168 countries and territories. www.em-

consulte.com/en/article/276237 is an academic article talking about “Information and therapeutic education of diabetic patients in French hospitals: The OBSIDIA survey”. www.lewishamandgreenwich.nhs.uk is the information website of Lewisham and Greenwich National Health Service (NHS) Trust which was established on 1 October 2013. This Trust is responsible for University Hospital Lewisham and Queen Elizabeth Hospital. www.uclh.nhs.uk is the information website of University College London Hospitals (UCLH). UCLH comprises the following hospitals: University College Hospital; Macmillan Cancer Centre; Elizbeth Garrett Anderson Wing; Hospital for Tropical Diseases; Institute of Sport, Exercise and Health; University College Hospital at Westmoreland Street (formerly the Heart Hospital); Royal National Throat, Nose and Ear Hospital; Royal London Hospital for Integrated Medicine; National Hospital for Neurology and Neurosurgery; Eastman Dental Hospital. www.worlddiabetesfoundation.org is the World Diabetes Foundation (WDF) official website. WDF is an independent, non-profit foundation based in Bagsværd, Denmark. It is one of the few funding mechanisms dedicated to preventing and treating diabetes in developing countries. https://www.ncbi.nlm.nih.gov/books/NBK343389/ is another academic article titled “Diabetes (type 1 and type 2) in Children and Young People: Diagnosis and management.”https://onlinelibrary.wiley.com/doi/abs/10.1111/imj.13649 is again an academic article titled “Access to a youth-specific service for young adults with type 1 diabetes mellitus is associated with decreased hospital length of stay for diabetic ketoacidosis”; https://www.joslin.org is the Joslin Diabetes Centre dedicated to conquering diabetes in all of its forms. It is the global leader in diabetes research, care and education. https://www.royalfree.nhs.uk is a pioneering organization and playing a leading role in the care of patients to provide world class expertise and local care.

As we can see, the results are related research Institutions, world-leading diabetes organizations, academic articles, diabetes foundations and some hospitals but mainly located in UK.

Therefore, finding appropriate pages through a search engine relying on web contents or makes use of hyperlink information is very difficult. Especially, it relies too much on how people search the information. Do people only type the keywords or type the whole sentences? Which

keywords can be chosen to precisely describe the questions? Do people get the bias information because of the location they are? With the increasing importance of the Web for an ever-broader spectrum of human actives, the explosive development of the information volume of the website has caused users to be unable to locate the information. How to effectively extract and utilize the online information is becoming a real challenge.

Every day, we are being faced with a variety of situations involving children and adults with type 1 or type 2 diabetes, gestational diabetic patients, healthcare professionals, parents who care for their diabetic kids, or companies who hire people living with diabetes or among others, where the needs for care are way more complex to be met or integrated solutions might not even be available at the moment.

In China, ailing people often take counterfeit medications and as a result their conditions are getting worse and some of them even develop life-threatening complications because they got the incorrect information online. A father, lacking of adequate knowledge and then gets misinformed and unprofessional medical advice online, asked his daughter who was just diagnosed as a type 1 diabetes, to take alternative drugs made of Chinese herbals instead of insulins, hoping that this alternative medication could cure his daughter because he read it online. Obviously, the Chinese herbals did not cure his daughter's diabetes. Instead, due to the steep shortage of insulin rationing, the girl required urgent hospitalization since she was suffering from a severe acute complication called Diabetes ketoacidosis (DKA) and her doctor even issued the "critically ill" notice. Thanks to the doctor's efforts, she survived the near-death episode.

Another tragic example involves a young professional who has diabetes and failed to land a job because the employer tried to assess the potential risk of hiring a diabetic by searching information online. Of course, he found out several negative aspects and led to the conclusion that diabetic people couldn't perform as well as healthy ones at work in general. Without a job, no income or insurance, that young man was nearly devastated and almost went blind because he could not afford the medications to properly manage his diabetes. In the end, he committed suicide.

We can go on and on with more tragic and horrific cases. What we learn from those cases is that the incorrect or wrong information on diabetes can have terrible consequences for our

diabetes community. The fact that a plenty of websites still offer disease information and medical advice without supervision or monitoring keeps us awake at night and also make us wonder how we could help the people in need to navigate the objective unbiased diabetes information online.

In summary, current search engines have certain limitations, such as:

- Users in different situations and different backgrounds often have different search terms instead of only few key words. And the requirements that the general search engine returns results in many pages that users do not care about.
- The goal of the current search engine is to maximize the network coverage. The contradiction between limited search engine server resources and unlimited network data resources will be further deepened.
- The richness of the World Wide Web data format and the continuous development of network technology, such as a large number of pictures, databases, audio/video multimedia, etc., a general search engine often has no power for these data-intensive and structured data, and cannot be very good to discover and get.
- Most search engines provide keyword-based retrieval, and it is difficult to support queries based on semantic information.

At the same time, even with some specific queries, a search engine doesn't give the accurate webpages. For example, if we try to find the top 10 blogs about diabetes using Google, when we type "top 10 best blogs about diabetes", it only offers the list of blogs instead of giving the specific blogs themselves. (see figure 5) Due to the complexity and variety of information on the website, how to integrate and display various information to the users is a difficult problem at present.

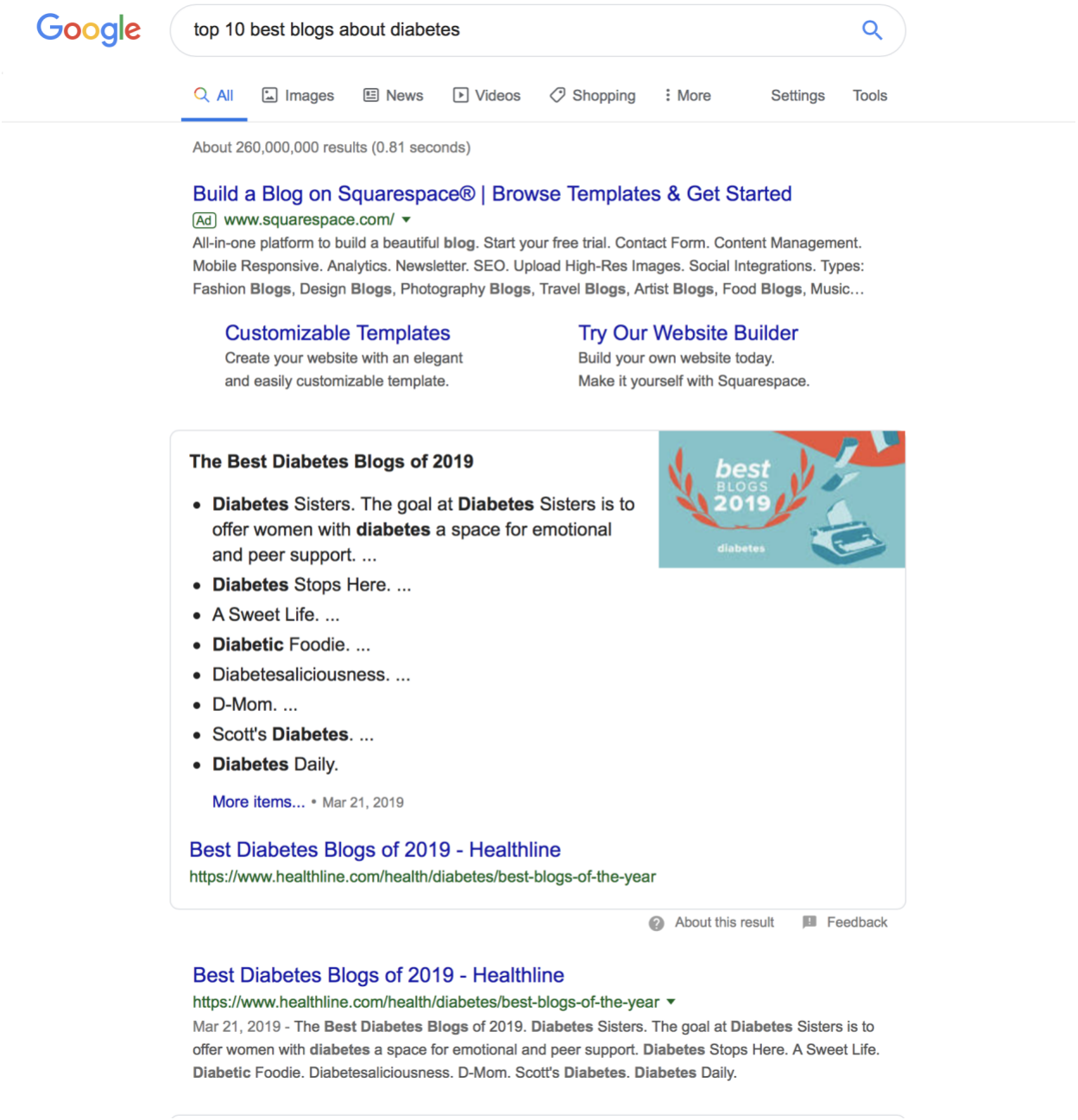


Figure 5 The website which offers the best Diabetes Blogs showing by Google.

1.4 Hypothesis

The World Wide Web (WWW) is an enormous virtual network of Web pages connected by hyperlinks. It represents just one of many examples of the topology of complex networks. Complex networks describe a wide range of systems in nature and society. While traditionally

these systems have been modeled as random graphs, it is increasingly recognized that the topology and evolution of real networks are governed by strong organizational principles.

Our hypothesis is that the web offers a global view that can be visualized and used to find relevant information. All stakeholders involved in diabetes have connections of some sort that can be revealed on the WWW as an organized world of communities rather than a randomly organized network.

1.5 Objectives

The presented PhD work has four main objectives:

- to provide a replicable practical methodology to visualize the diabetes network of websites to make sure all key health web players have connections with one and another.
- to help people get un-bias information by visualizing the network of diabetes in the digital world.
- to use the networks of World Wide Web to highlight the connections of some stakeholders involved in diabetes as an organized space filled with communities rather than a randomly organized network.
- to define some scenarios using the network to demonstrate its interest.

In this work, we propose to map the hyperlink structure of diabetes stakeholders online. As a matter of fact, the web contains crucial aspects of the embodiment of social factors: personal blogs, institutional websites, health-oriented media, etc. To address the relationship with stakeholders in the diabetes area, we have examined how classical studies of network explorations can be brought to use in the context of diabetes to answer the following questions: Who is connected to whom by which means? Which organizations receive support from which ones? What resources or information are published through which platforms? What are the relationships between charities and governmental agencies? How individuals, companies and organizations interact together? What is the ecosphere of the diabetes world?

A corollary expected impact is that this map can help the diabetics to access useful information regarding the neighboring context (such as information on diet or physical activity)

that search engines don't display at first when looking for information about diabetes. At the same time, this can also help pharmaceutical companies and Non-Government Organizations to find influencers or key opinion leaders and help governmental agencies to take effective actions that reflect more efficiently the needs of diabetes communities. The visual representation will be helpful to quickly and effectively identify areas on the graph where improvements can be made.

To achieve our objectives, we worked on a methodology which is made possible by combining a range of tools to assist the creation of websites networks maps on diabetes. In our research, web crawler tool and networking visualization tool will be combined to create the map of diabetes-related websites and semantically interpreter it.

1.6 Presentation of the manuscript

This manuscript contains 6 parts. In Chapter 1, we mainly introduce the general situation about diabetes nowadays and with the recent dramatic rise of diabetes occurrences raises, several questions that need to be addressed: how do people react after they are diagnosed diabetes? What will they do for long term self-management? Where can they find relevant information and how easy is it to access it? How can they get psychological support and community help? We notice with the current search engines that the specific information is not easily found by people who just knew few about diabetes. As a result of this work, we propose a map-like information to help people access useful information from diabetes web networks.

In Chapter 2, we review the hyperlink network studies applied in other different disciplines and also the network visualization, especially with map-like information visualization approach. As a state of arts part, we also present the three approaches for semantically interpreting the map-like information visualization which is including ontology, terminology and natural language processing.

In Chapter 3, we introduce the state-of-arts tools we used to crawl (Hyphe) and visualize (Gephi) topic-sensitive networks. These tools not only offer a map with diabetes-related websites but also change the way to present the top-sensitive networks. At the same time, we describe the methodology in detail from the preparation of initial websites to visualization of the network

structure. Each step is accompanied by the decision criteria as to provide a replicable practical methodology for visualization of the diabetes network of websites. In the end, we present the final map (DiaMap) using Hyphe to extract 430 specific diabetes-related websites and Gephi to visualize with force atlas 2 layout.

In Chapter 4, we use a semantic approach to explain the 5 clusters detected by applying the community detection algorithm for detecting the similarly connected nodes. The purpose is to better understand the common interest shared by the same clusters. To achieve that, we divide this exploratory study into three steps: (1) to define and organize the best tags describing the semantic content of websites; (2) to define and set an annotation process to annotate the 430 websites manually using the tags; (3) to apply various machine learning methods to predict the cluster of a website from the annotations. We present 7 different machine learning models we used to predict the clusters and also list the top 10 tags contributing the most for predicting the clusters.

In Chapter 5, we try to access DiaMap for accurate information retrieval by comparing with the current popular search engines. The principle is to choose 5 questions on diabetes of interest to a real user and to search for the corresponding websites on the search engines and in an equivalent manner, using the corresponding tags on DiaMap. The comparison will be made between the websites proposed by the search engines and the websites proposed by DiaMap according to a certain protocol.

After the comparison, we found that search engines can offer a wide range of websites to users to answer their questions. However, users need to read each website and still have to judge whether the websites are reliable by themselves. DiaMap is more like content organized by online communities on diabetes. With this tool, we help people find more precise and topic-focused websites, using tags and clusters to distinguish them.

In Chapter 6, the final discussion and conclusions point towards an informal research agenda for further work in that area. DiaMap presents the visualization of map-like information from diabetes-related websites to manifest online diabetes-related content and its structure. Different from traditional search engines, DiaMap presents all the communities and uses tags to

identify the most relevant and reliable websites. It somehow changes the way people navigate online information about diabetes and could be an alternative to increase the current search engines. DiaMap can become a search engine that focuses on the global online world of diabetes to help people with their diabetes.

In the future, we will increase the size of the dataset to more than 5000 websites in DiaMap. Our idea is to use automatic natural language processing to learn the most relevant tags from websites and automatically annotate a much larger corpus. This manuscript concludes that the online space on diabetes exists and that this new virtual world as an open resource should be known to everyone in its entirety. This is just the entry point to introduce how to combine network visualization and both topological and semantic community detection to do network analysis for diabetes. This new approach can also be extended to network analysis for other chronic diseases helping patients and their families to navigate in a traditionally search engine domain digital world.

2 Chapter 2 State of the Arts: Hypertext Network Analysis

When confronted to a chronic disease such as diabetes, regardless if it is as a patient, a caretaker or a concerned citizen, active information-seeking users have two major sources of information: the web and healthcare professionals (Boukacem-Zeghmouri & Schöpfel, 2013)(Kuske et al., 2017). While the latter is recognized for its authoritative status, studies show that most users turn to an online source of information to retrieve useful advices, social support, read stories, buy medical devices or pharmaceutical products, propose oneself help to organizations and so on. To navigate on the Web, one uses a web browser and combination of tools to find and display the web pages, organized in web sites, created by both algorithms and humans.

When seen through the prism of a web browser and selected websites such as general search engines (NW et al., 2012) the web appears to be a collection of disconnected resources dynamically curated by search engines black box algorithms (Pfaender et al., 2006). But the web itself is more than an unordered collection of online document, it is a non-Euclidean space with a distinctive structure (Kleinberg & Lawrence, 2001). This structure is used by search engines and users alike to rank information and navigate through the online space to discover further resources. As a consequence, deciphering the Web space and how it exposes information allows to: (a) better understand how users contextualize and build representation of the topics at hand, which in our case is diabetes; (b) how the topic itself is organized as a sub-space of the Web with its own characteristics.

To study the Web space of diabetes, our strategy is to understand the Web fundamental properties; then make use of an existing set of multidisciplinary methodologies and theories to analyze its structure and content; all the while using data visualization to assert our hypothesis and provide new insights.

2.1 World Wide Web

The study presented in this PhD dissertation uses the Web as the primary fieldwork. For that reason, it is crucial not only to comprehend the Web properties but also how these properties can affect subsequent methodologies we will formulate. To achieve such an objective, we started from the Web substrate, to proceed with the Web definition including the semantic or Web 2.0 initiative.

2.1.1 Internet / Web

Internet is an interconnection of networks of connected equipment exchanging bits of information using a variety of protocols (ISI, USC, 1981, >> Information Sciences Institute, University of Southern California (1981) Internet protocol, Published as Internet Requests for Comments. Retrieved March 17, 2020, from https://tools.ietf.org/rfc/rfc791.txt). The highest level of hypertextual protocol is HyperText Transfer Protocol (HTTP) and is define as a protocol to exchange hypertextual content between a client and a server. There is an equivalent protocol with end to end encoding: HTTPS (Nottingham, 2019 >> Nottingham (2019) Well-Known Uniform Resource Identifiers (URIs), Published as Internet Requests for Comments. Retrieved March 17, 2020, from https://tools.ietf.org/rfc/rfc8615.txt).

The server and the clients both have a physical location, that is a physical computer connected to the network on both ends allowing a browser (client side) to ask content to a server (server side) and eventually getting a response back. Internet is a global set of backbones with IT equipments transferring digital data and supporting several protocols among which **http** and **https** are but a special and high-level case. There are important consequences of internet being supported by physical IT equipment for research on online communities. As the Internet is embodied by physical devices, the devices are in a place belonging to an administrative body whose rules extend to data being transferred on its premises (Deibert et al., 2010). Therefore, the content being transferred from user to the server can be subject to censorship or arbitrary changes by said administrative body which, affects the very nature of any scientific investigation on the web and the communities it sustains. Some pages might not be accessed while some pages might be changed based solely on user or server locations in a sovereign country. Therefore, to replicate experiments

and study of an online space, one needs to also replicate the condition in which the online content was retrieved with an emphasis on the location.

Moreover, as user's location can be estimated by looking at the location of the equipments the information travels through. Some web services use this information to serve a location specific content to its users. This is often employed by search engines to provide specific content to users and better match their potential expectations (Kliman-Silver et al., 2015). But again, the consequence for research are non-negligible as the same online inquiry about a diabetes website might be different based on the end user (or client) location.

To address these issues, we chose first and foremost to study diabetes online community from an international point of view to avoid potential censorship and get the most largely shared information where location would not be critical. Any country specific or language specific investigations where location matters most were only added in a second time once the core methodology was identified and proven. Furthermore, the result of this manuscript must be considered from emanating from INSERM lab in Paris which was the physical location of the computer handling the data harvesting.

2.1.2 Network of Hyper Textual Documents

On top of the internet, the World Wide Web (*WWW* or web), commonly known as the Web, is a collection of information which is accessed via the Internet. Although researchers somehow conceptualized the Internet differently, it was originally described as the network of networks (Rob Kling, 2006). The basic structural elements of the Internet are hyperlinks. Hyperlinks can be defined as technical capabilities that enable a website (or web page) to be linked to another website (or web page). One hyperlink points to a whole document or to a specific anchor within a document. The text containing a hyperlink is known as a hypertext (KLEINBERG, n.d.) as it adds a new dimension to the text. A user following hyperlinks is said to navigate or browse the hypertextual network or the web. For example, in an online reference website such as Wikipedia.com or Google.com, many words and terms in the text are hyperlinked to definitions of those terms. In this type of hypertext, hyperlinks are often used to implement reference

mechanisms such as tables of contents, footnotes, bibliographies, indexes, letters and glossaries already existing in traditional text and amplified with interactivity.

But they can also add a secondary content represented by the hyperlinked texts, therefore changing the very nature of the first text now in the center of a web of hypertexts. In some rare cases hyperlinks can be bidirectional: they can be followed in two directions, so both hypertexts act as anchors and as targets. The outcome of following a hyperlink depends: on the nature of the hyperlink (if the hyperlink indicates a specific protocol, *mailto* for example, it will launch the default mail software); on the user browser (programmed to follow certain rules when hyperlink is activated); on the server, itself hosting the text (the server can decide to serve pages differently depending on a set of parameters including user's navigation history and actual behavior). For instance, on the World Wide Web, generally, most hyperlinks cause the target document to replace the document being displayed, but some are marked to cause the target document to open in a new window or start an additional scripted action.

Since the Web was invented at the beginning of the 1990's, it has gradually become a major media platform itself hosting billions of resources accessible from a Uniform Resource Locator (URL). It also hosts dozens of sub-media in the form of specific websites proposing social media (like Facebook or Twitter), scientific knowledge or any other niche media (Réka Albert et al., 1999).

The Web rapidly became a source of scientific investigation in itself and the network structure of a hyperlinked environment can be a rich source of information for interdisciplinary studies, as long as means to collect, analyze and understand it are met (Jon M. Kleinberg et al., 1999). Hyperlinks allow individuals or organizations creating webpages on websites to extend their *social* relations by simply and directly connecting people or groups anywhere in the world. While the nature of the link is unspecified by the medium, creating a link to another hyper textual resource is seen as a strong intention to reference, denounce, qualify, affirm opinions, knowledge, and by extension authors behind them. In a hyperlinked system, the individuals and organizations can be linked together, exchange information and maintain partnerships using hyperlinks to

associate webpages and create a common background, interest or project (M. Newman et al., 2011).

Moreover, through hyperlinks, individual or organizations websites' owners become actors with a potential to influence that can directly affect other website's trust, prestige, authority, or credibility (Park, 2003). If considered in large exception, hyperlinks are a key primary structure supporting networks of knowledge, people, organizations, or nation-states. Thus, studying the hyperlink structure, we can reveal or infer the social structure and/or communication structure among those groups.

As of recently, more and more scholars tend to consider websites as actors themselves, regardless of their actual ownership. From this perspective, an actor is a hybrid element belonging to a person, private company, public organization, city, or nation-state but whose identity is inexorably merged with the website semantics and hyperlinking strategy. These websites, or nodes, are linked through their hyperlinks. Despite the relatively short-lived existence of the web, as the structure of hyperlinked networks continues to change, the role of the web as media for communications has to be acknowledge. Patterns of hyperlinks designed or modified by individuals or organizations owning the website, and link brought through social features such as comments not only reflect the owner's communicative choices, agenda or goals but a larger hybrid complex social construct (Jackson, 1997). As a result, a surprising growth of interdisciplinary hyperlink network studies have been witnessed across many disciplines, principally social science, in the past two decades (Park, n.d.).

Before going further with web analysis, we must clarify the definitions of **webpage**, **website**, and **web entity** in the scope of this thesis. While these terms might appear rather self-explanatory, they hide a complexity due to the very nature of the web:

Webpage – a webpage is the document displaying in the web browser corresponding to one URL. This document is always a construction aggregating multiple resources into a coherent visual interactive experience. The HTML format is dedicated to the structure of the page and usually hosts the content. The HTML itself is rendered visually with as stylesheet often put in a

separate file as a Cascading StyleSheet or CSS. Interaction is handled with scripting languages most often javascript and in modern web it comes in multiples separated files representing data and libraries. All multimedia content is also added to the page by the web browser using internal and/or external resources. A video from the popular website Youtube for example, displayed in a random blog webpage comes directly from outside of the blog itself. To add to the complexity, a page can also embed other pages directly inside itself as a kind of window to another webpage called Iframe. As a consequence, the webpage itself, as it is experienced by a user, has no technical existence. It is a mix of resources having a unique and temporary existence when displayed. Furthermore, an automatic analysis of text content coming from the sole HTML is rather limited compared to what the user may experience. Therefore, future experiment in this thesis took this into account and largely relied on time consuming but very effective human analysis of webpages. It has to be noted that the webpage User eXperience (UX) also varies with web browsers as they interpret the page sometimes differently; so, there is no guarantee that a page seen at a time and place by a user will be the same by another user at the same time/place. For research perspective, we chose a mainstream up-to-date web browser (Chrome) to guarantee that the experiment will be as close as possible to what most users' experiences.

Website – a website is a collection of webpages offering a coherent experience as a whole. The coherence is a by-product of all or part of the following, non-exhaustive components: semantic based (similar topic, diabetes blogs for example), author based (similar author, company, organization or brand, microsoft.com for example), UX based (similar experience through the pages, twitter.com for example). Most often the website is identified by a top-level domain name (TLD) like inserm.fr, that is registered with a registrar authority and point to a server on the internet. But a consistent domain name across several webpages do not to guarantee a website coherence or integrity. For example, platforms like hypotheses.org hosts thousands of websites under its domain name. The websites receive a subdomain name like datafabrik.hypotheses.org. In this case the subdomain identifies the websites. But it does not stop here as some organization can host a platform far right the URL path. So, the definition of what a website is remains under the user's appreciation of a perceived coherence of a collection of pages. The domain name is a welcome mnemonic but shall not be the reference for a website definition.

Web entity – to study domain specific network of web resources we decided to rely on a more flexible definition of a website or webpage that the one defined by the URL only that acknowledge website perceived coherence and that is, as seen above, improperly representing the complexity of our resource. A web entity is an URL schema (a fragment of the URL path) representing a coherent ensemble of webpages. The web entity is the root of the tree represented by the URL schema after which all webpages are considered. If the web entity is a single endpoint page, then the entity is solely represented by the page. On the other hand, if the web entity is a TLD, then all the pages of this domain are considered relevant for the entity. The entity allows to have a fine control over what is considered or not as a node of the topic network under study (Jacomy et al., n.d.). It allows to add media platform like twitter by considering only relevant accounts or even specific posts made on the microblog publication website. The entity can be automatically associated to know platforms but ultimately, the web researchers shall define it manually according to their user experience.

2.1.3 Semantic Web

One of the major drawbacks of the web, is the inability to qualify the hyperlink with the intention it carries. The hyperlink is just a link between two hypertexts with no explanation of the nature of the link.

One way to represent effectively the meaning behind links is to add a meta-information to the link itself and explicit it. As a matter of fact, this initiative was proposed rapidly after the first definition of the web by CERN inventors in 1994 (Shadbolt et al., 2006). Coined *Semantic Web*, this initiative was also dedicated to create a Web of Data even though the term semantic remains connoted with a linguistic reference (Luís Miguel Oliveira Machado et al., 2019). Both acceptation of Semantic Web remains the same for the objective of the present thesis: it proposes to add context to individual fields inside the hypertext. The context is provided though a Resource Description Framework (RDF) and assign a URI (Universal Resource Identifier) descriptor to the two ends of a hyperlink and to the relationship between them, thus making a triplet. RDF are triplets' description of webpages and their relations. It adds an extra layer of information, very useful in theory, but it never received the drive it needed to become widely adopted from web industry and web creators.

The reasons for this lack of interest from the general public are multiple. The simplest one is that creating triplets for every piece of information inside a hypertext involves an extra work to make explicit a previously implicit contextualization. Most often, authors are not ready to do it nor have the proper tools to engage in such a work. Moreover, the syntax is far from simple and the RDF have to be properly written to be used by semantic web engines. Therefore, semantic web is mostly dedicated for internal use in knowledge databases or in very specific areas of the web handled by web experts, none of which are expected in diabetes web networks. Finally, the semantic web lies in the trust one can put in the description. If one would perjure oneself when describing resources, it would corrupt the whole chain of trust the hyper-network lies upon.

Unfortunately, such a behavior is common to get better exposure on the Web. As the web fame resorts, largely to general search engines, for one web page to be found, optimizing one position or rank in search engine has become a common skill for web media strategists. Called Search Engine Optimization (SEO), most of the techniques are admitted (usually referred to as **white hat** techniques in reference to benevolent computer science experts) but quite a few of them are not considered as acceptable (referred to as **black hat** or malevolent computer science experts) and are actively monitored by search engines banning the offenders (Gaharwar & Shah, 2018). The black hat techniques rely on disrupting the chain of trust one can put in hyperlink by artificially creating hyperlinks (farm linking), falsely associating certain keyword to certain web pages (keyword stuffing), serving different content to users and to search engines visiting the page (cloaking), etc. The Semantic Web would be extremely sensitive to these contexts alteration as it still relies mostly on author good will with no control authority. The search engines are **de facto** the control authority. Only a change of paradigm would make it relevant in the web in general.

As a consequence, we cannot rely on semantic web for our purpose of studying Web networks of diabetes even though semantic web would fit our purpose perfectly by allowing the hyperlinks to be explicit on their intention allowing us to understand the diabetes web in depth. Besides, semantic web is not largely used while we plan to explore the diabetes space a common Web user would encounter it. Another corollary consequence is that we need to beware of black hat techniques when analyzing network of web pages and web sites. Fortunately, these can be

avoided or discovered when using a combination of network analysis and human screening of pages though visualization and web browser detailed examination.

2.2 Network Analysis

A network of web entities is but a particular instance of a more general complex network. Network analysis is a scientific field in expansion and complex network analysis is itself a network that can be studied through a set of properties and algorithms ("ECT volume 26 issue 5 Cover and Back matter," 2010). Complex network analysis holds not less than 11 areas of applications (Luciano da Fontoura Costa et al., 2011) from protein exchange to social networks, from road networks to web networks. These networks have specific statistical properties (Reka Albert & Barabasi, 2002), largely studied (M. E. J. Newman & Girvan, 2004) (Boccaletti et al., 2006) and providing numerous examples (Alhajj & Rokne, 2014) (Ed Bullmore & O. Sporns, 2009).

We will focus our thesis on a web network but our first concern is to make sure the software we use to build our web network on a specific domain will effectively provide an uncorrupted web network, that is a complex network with 3 main properties whose range for the properties is within the ones of web networks already studied.

Then, as we focus on a web network representing a social organization for diabetes, we will make use of Social Network Analysis (SNA) that we will be detailed in this section. Lastly, as we plan to perform as good as search engines or better for diabetes related topics, we will present basic methodologies used by search engines to provide adequate content to users.

2.2.1 Complex Network

First and foremost, we need to prove our network not only belongs to complex network but is a proper web network in order to be able to use the field algorithms and methodologies. Complex networks are characterized by three main properties: the *average degree distribution*, the *clustering coefficient* and the *average path length*. Each of these properties is linked to the nature of the network.

2.2.1.1 Degree Distribution

The degree of a vertex (i.e. node) in an undirected graph is the number of edges (i.e. connections or links) adjacent to this vertex. The degree is calculated for each vertex in the graph to get the vertices distribution of degrees in the graph. In a directed graph, this leads to 3 measures: an in-degree distribution for edges leading to vertices; an out-degree distribution for edges leaving the vertices; and a total degree distribution, sum of all in and out edges.

Some graph has a small number of vertices with an important number of edges whereas most of vertices have a small number of edges. The nodes with more connections are called hubs as they usually lead to hundreds if not thousands of neighbors. This difference in degree distribution makes the graph appears as if it has no scale. In this case it is called a scale free network and hold important characteristics such as being robust if a random vertex disappears but being vulnerable if the hubs are suddenly removed (BARABÁSI & BONABEAU, 2003).

Web networks are typically scaled free networks (Barabási et al., 2000). Degree distribution in a scale free network shall have a lot of lesser degree vertices and few larger degree vertices. Hence, they usually follow a power law degree distribution whose slope can be measured and compared to similar networks in the literature (Reka Albert & Barabasi, 2002).

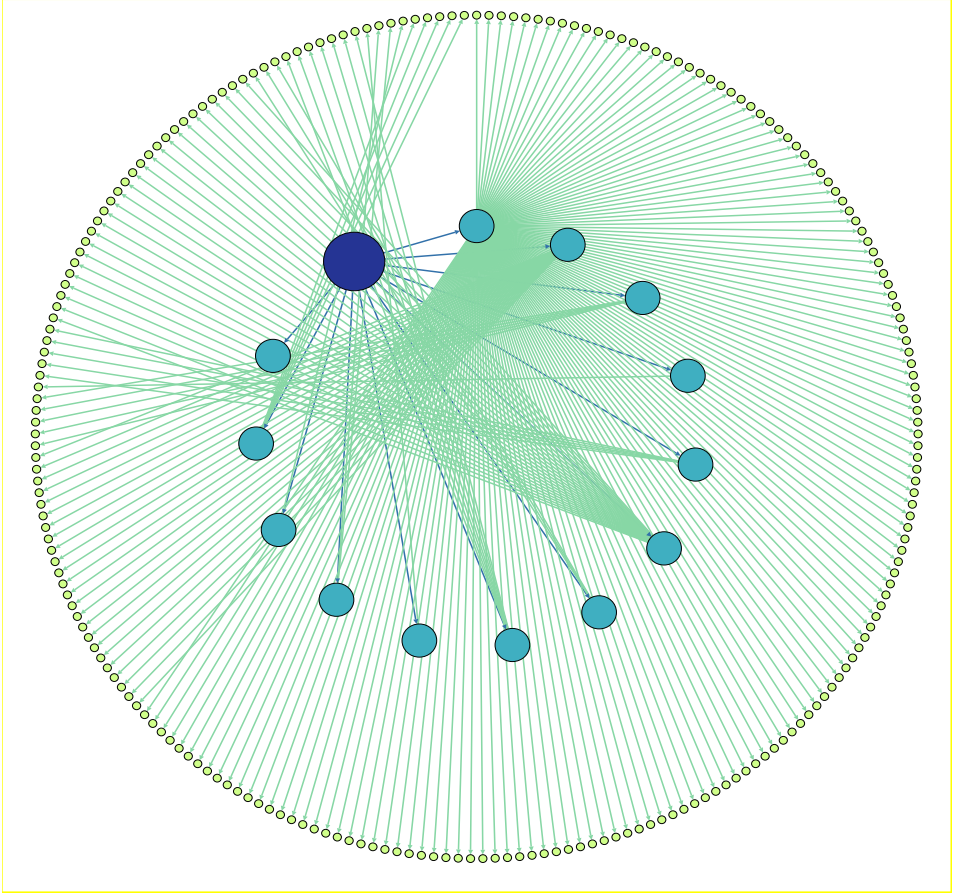


Figure 6 Power law distribution of hyperlinks starting from Diabetesramblings.com, depth (1) = 13 nodes, depth (2) = 225 nodes

We can see this effect in the visualization of the web of figure 6. Starting from one website for our corpus (Diabetesramblings.com), we can find 13 direct neighbors and 225 neighbors of these neighbors with an exponential increase with each step. We used a python programming library to test the power law fitting of the degree distribution of the corpus and estimate different fit (Alstott et al., 2014) in chapter 4. A web network typical coefficient for a power law define as $degree(x) = \alpha x^{-\alpha}$ is $\alpha_{out} = 2.45$ and $\alpha_{in} = 2.1$ (Réka Albert et al., 1999).

2.2.1.2 Clustering Coefficient

The second property of the complex network that we intend to use is the average clustering coefficient. In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together (Holland & Leinhardt, 1971). It is calculated in two steps: (1)

computing the local clustering of a vertex defined as the fraction of actual triangles for this vertex over all the potential triangles in the neighborhood; and (2) averaging the local clustering for the whole graph (Easley & Kleinberg, n.d.).

In most real-world networks, and in particular social networks or web networks, nodes tend to create tightly knit groups having a high density of ties. This is especially true if this probability is greater than the ties found in a random network of the same characteristics (e.g. same vertices count and edges count and link probability) (DJ Watts & SH Strogatz, 1998). Typical value found in social networks (coauthor-ship for example) yields a clustering coefficient comprise between 0.3 and 0.8 while the random network equivalent is below 0.001 (M. E. J. Newman, 2001). To calculate this value in our graph and a random equivalent, we used the python library NetworkX (Hagberg et al., 2008).

2.2.1.3 Average Path Length

The third property we need to look at is the average path length or more accurately the average shortest path length. This is a measure of the shortest path between all pairs of vertex in the graph which is then averaged. The shortest path itself is the number of edges along the shortest path between a source vertex and a target vertex. It corresponds to a diameter that can grow quite large in a network of unrelated vertices while being small in networks where vertices are more closely related. While the measure is quite simple in its spirit, its computation can be rather computer intensive if performed on large graphs as every pair of vertices needs to be examined and a path sought between the source and the target. Typical values for a web graph with a social dynamic should be close to 3.5 while a network created from random webpages on the web can make the average path length grow to 11 to 12 (Reka Albert & Barabasi, 2002).

When the average path length is relatively small and the clustering coefficient high, the network is said to host a small world phenomenon (DJ Watts & SH Strogatz, 1998) in reference to the small world proposed by psychologist Milgram (Blashfield & Aldenderfer, 1988) where distance between random people in US is said to have a small shortest path of acquaintance of 6.

These 3 properties are largely used to describe complex network. They are useful to verify that network belong to a known family following a set of principles. Among the families of network, the network of a topic such as diabetes is expected to belong is the social network family.

2.3 Social Network

A social network is a network of individuals sharing a bond of a particular nature (Scott, 1988). The very nature of their bond does not need to be physical or visible as long as the ties can be captured. As mentioned before, websites themselves are considered as actors, assimilated to individuals, with a hyperlink acting as a materialization of their bonds or ties in the digital world. This allows us to use social network theory, indicators and properties to measure the network of web entities as an intrinsic social organization.

The measurements include but are not limited to: modularity, degree centrality, closeness centrality and betweenness centrality as the most common ones. To illustrate their differences, we calculated them on a small sample of our diabetes network that you can find in figure 7.

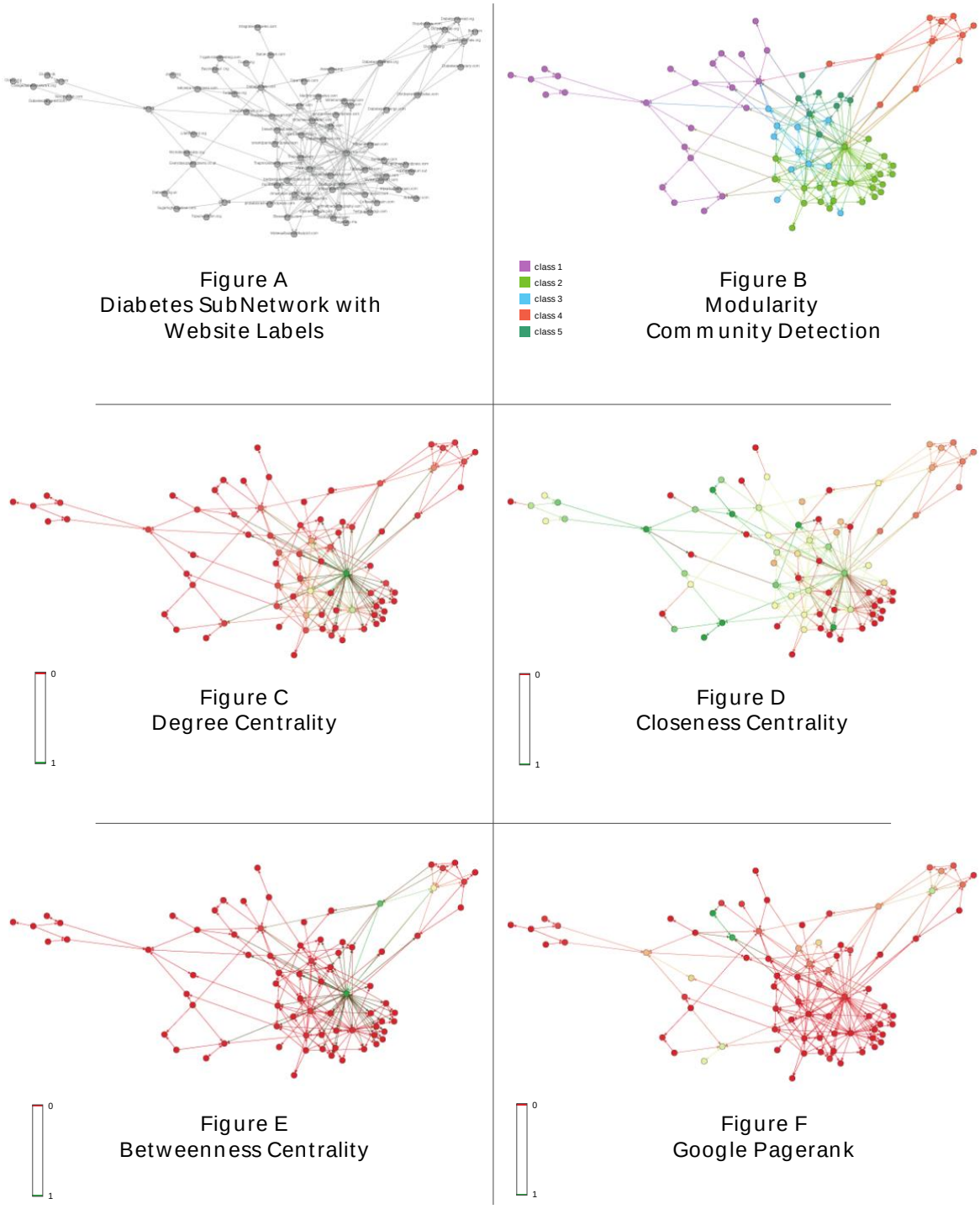


Figure 7 A is the network itself with the name of the vertex as found on the web. Subsequent figures are a visualization of the different centrality measures applied to this sample.

2.3.1 Community Detection (Modularity)

To analyze a network structure beyond its general properties such as the average path length or clustering, one often tries to classify or divide nodes into modules or communities. Such communities are sought to be stable in term of density inside the community but different from the density outside of the community. Hence, community detection algorithm aimed at detecting clusters of similarly connected nodes (Clauset et al., 2004) *(Blondel et al., 2008)*. In order to do so, they use a measure of modularity of the different partitions of a graph. Modularity is defined as a scalar value that measures the density of edges inside the partition as compared to the density of link between partitions. Different partitions are created and tested with the goal of refining and optimizing modularity incrementally. An optimized modularity yields well defined density communities if the network allows it (M. Chen et al., 2014).

In the case of social network, we can expect a community of individuals to have a similar habit of connecting together, that is a similar density. That is the reason why this measure is widely used in social network to find groups of individuals forming communities. Communities in network analysis then can assimilated to a social community. The result of community detection on a network is shown on figure 7 B with one color for each community. 5 distinct partitions were found in this case and it gets a rather convincing visual confirmation.

This measure has a limited relevancy if the network is heterogeneous. In a too diverse network, some communities might indeed have a similar modularity but some will be very different in which case modularity tends to mixed them up or divide them arbitrary rendering the communities useless. This is especially true for scale free networks where degrees are known to vary a lot (see degree distribution above). Web network are at the frontier of both world where their social networking characteristic is favorable for community detection but their scale free nature might endanger it. It might very well fail depending on the web network at hand. Most of the time, the network is the home of both phenomenon. New algorithms have been proposed to compensate for this sensibility to heterogeneity such as stochastic block model (Karrer & Newman, 2011) or nested stochastic block model (Peixoto, 2014), but these models went beyond the scope of this thesis and were not explored as of yet. However, they represent a very interesting

potential future work in correlation with the language tools that will be described later in this chapter.

2.3.2 Degree Centrality

The second useful measure in social network analysis is the degree centrality. The degree is the number of edges connected to a vertex. The degree centrality is the fraction of vertices connected to a specific vertex. The values are usually normalized by dividing the maximum degree in the network n-1 where n is the number of vertices in the network.

If in a social network, the node represents a person and the edge represents the relationship, then the biggest degree centrality of a node is, the more friends there are. A person with a big degree can be a celebrity. If you need to spread some news, such a person is the most suitable, because he or she plays the role of key opinion leader (KOL) (Srinivas & Velusamy, 2015). In figure 7 C, we use color to represent the degree, green represents having the most links and the red represents having less links. For example, the vertices *textingmypancreas.com*, *diabeticscornerbooth.com* have more websites directly connected to them. This property makes it easy to find very connected individuals, popular individuals, individuals who are likely to hold most information or individuals who can quickly connect with the wider network. Degree centrality is the simplest measure of node connectivity. Sometimes it's useful to look at in-degree (number of inbound links) and out-degree (number of outbound links) as distinct measures, for example when looking at transactional data or account activity.

2.3.3 Closeness Centrality

Closeness Centrality is the third measure used in social network analysis. The closeness centrality of a vertex *u* is the reciprocal of the average shortest path distance from each other vertex to *u* over all *n-1* reachable nodes in the network of *n* vertex. (Sabidussi, 1966). In other work, closeness is high if a vertex can connect to many other vertices easily. A higher value of closeness indicates a higher centrality.

In a social network, the bigger closeness centrality of a person indicates that this person can quickly reach all people. This person may not know lots of people, but his or her friends can be well-known. Hence, this property is to find the individuals who are best placed to influence the entire network most quickly. Closeness centrality can help find good *broadcasters*, but in a highly connected network, you will often find that all nodes have a similar score. In a scale free network, it would lead to the hubs. What may be more useful is using Closeness to find influencers within a single cluster. Figure 7 D shows the Closeness Centrality of the diabetes-related websites. The left part has a rather closeness centrality because closeness is calculated inward and not outward in a directed graph. In the figure, we end up easily onto the left because these websites are authorities, meaning that they get a lot of inward traffic but few outward traffic.

2.3.4 Betweenness Centrality

Betweenness centrality is the fourth measure in social network analysis. The betweenness of a vertex *u* is the sum of the fraction defined as all-pairs shortest paths that pass-through *u* divided by all-pairs shortest paths (Brandes, 2001). Betweenness centrality measures the number of times a vertex lies on the shortest path between two other vertices. It does this by identifying all the shortest paths and then counting how many times each vertex falls on one. The betweenness centrality can be understood as the amount of resources a node has. In a social network, a high betweenness centrality of a vertex indicates that the vertex has more resources and the resources are irreplaceable (likely pass through him, there is no shortest path). Comparing with Closeness Centrality, betweenness centrality is more likely to find the Individuals who influence the flow around a system. It is useful for analyzing communication dynamics, but it should be used with care. A high betweenness centrality could indicate someone holds authority over, or controls the collaborations in a cluster of a network. It can also indicate the vertices serving as a bridge between communities as bridges are usually on the shortest paths. In figure 7 E, together with textingmypancreas.com, already highest degree node, diabetesstophere.org and diabetes.org have a high betweenness denoting a tendency to serve as a bridge. It can mean that they are a kind of authorities in this network.

One of the important outcomes of social network analysis is to identify a central node, in this case, a central Web site, generally defined as the site that provides the most and/or shortest

connections to other members within the group. The central web site usually plays the role of hub, and authoritative or prestigious site. Closeness centrality is used to determine which web site has the shortest path to all other in the group. Betweeness centrality refers to the frequency with which a web site falls between pairs of other sites in the group and represents the potential for control of communication, as a broker or a gatekeeper (Freeman, 1980). If added to degree centrality and modularity, we dispose of a set of measures to understand node importance in a network and by extension the role of actors in it.

2.3.5 Search Engines

Third, one objective of the present study is to improve the efficiency for the average web user to retrieve information on a specific diabetes related topic and get a contextualized knowledge of the place of the topic in the diabetes community.

As it was presented in the introduction, the most prevalent entry point for such online health information are general search engines (NW et al., 2012), capable of proposing supposedly relevant websites through results pages (SERPs) from a sentence or set of words referred to as a query. For a given query, hundreds of thousands of results are provided, paginated into SERPs. This represents an important opportunity for finding a relevant information but the potential is rarely actualized as users favor only the first 3 to 5 results of the first page and vastly disregard the rest (Kim et al., 2015) (Cutrell & Guan, 2007). (Höchstötter & Lewandowski, 2009).

To rank information according to query, search engines use a two-step process. The first step is to collect as much data as possible. This is done by getting to a hyper textual resource and following thoroughly all hyperlinks exactly as a user would do. As the web is a scale free network, the quantity of data increases exponentially at each step but it allows to discover mostly everything. While continuously harvesting hypertexts, the second step is to use the current index made of document referencing words and reverse it to word referencing documents (Brin & Page, 1998). So, for one word, the search engine can propose several documents. The question remains as how to rank the documents as there are multiple ones for any given query and only a very tenuous fraction will be proposed to users let alone seen by them. The strength of search engine resides in its corpus (the documents database) and the ranking algorithm. If current ranking algorithms are

black boxes property of the search engines, early papers revealed that they use network properties to rank document (Page et al., 1999).

PageRank, for one, use only the network probability of a node to be visited by a random surfer to get to a page. The higher the probability to get to a page by chance of following links, the higher the PageRank. Moreover, the algorithm is recursive so being linked to high PageRank pages, increases PageRank too. The hypothesis here is that a good resource will get a lot of good quality links and the network will self-evolve to respectively ban bad content and promote good content. PageRank simply reads the network to extract a ranking and use this to propose the best document for a given query.

The first ranking for google.com was not based on the quality of the content but rather on the quality of the network around it. Yahoo, for example, took an opposite direction and relied heavily on human categorized documents (Ceci & Malerba, 2007). This is impossible at large scale and ultimately the semantic approach of yahoo simply enriched the network (and very large scale capable) approach of google.com.

To achieve the goal of proposing users a better way to find resources for a chronic disease such as diabetes, we needed to compare our result with predicted search engines result such as Google. This can be estimated by measuring the PageRank of vertex in a network. As seen in figure 7 F, the result of the ranking ignores the social network analysis centralities and new websites such as *beyondtype1.org* and *tudiabetes.org* would be highly ranked if a diabetes query was to be run against our network.

2.4 Applications examples

This section reviews the prior research that conducted a network analysis within the topics of international communication, e-commerce, interpersonal communication, and inter-organizational communication.

2.4.1 International Communication

Halavais examined the role of geographic borders in cyberspace using the hyperlink pattern of websites (Park & Thelwall, 2003). Specifically, he took a sample of 4,000 websites and analyzed their external hyperlinks and determined the total percentage of hyperlinks from the sites to various countries. Domains, which did not contain their geographic locations (for example, .com or .edu), were checked against the registrar to determine the country of origin[3] .

2.4.2 E-commerce

Palmer used the hyperlink methods to examine e-commerce (Palmer et al., 2000). When purchasing a commodity online, a consumer's trust (or perceived credibility) of a website has been regarded as one of the most influential factors in transaction process. Based on this theory, they used the number of inward hyperlinks to a website as an indicator of trust in the firms. They obtained the data from *Alexa.com*. The results revealed that the number of incoming links was strongly related with the use and prominence of Trusted Third Parties and privacy statements which are regarded as another trust indicator.

Krebs' study of Amazon.com indirectly revealed the roles of hyperlinks in relation to homophile attribute among online consumers (Krebs, 2000). Amazon.com provides customers with information about who bought this book also bought these books. It has a hyperlink so that prospective customers can take a look at the hyperlinked books directly. Krebs argued that the fact that people with similar interest bought those books contributes to persuading prospective consumers to buy them. Choosing a specific book as a focal node, he built an *ego* and *alter* network between books. This enabled him to see how the hyperlinked books are interconnected and what position they occupy in the networks. Also, the books were clustered according to a topic and he analyzed the role of individual book within cluster and among clusters.

Park explained websites' hyperlink affiliation networks as a function of the credibility among websites and the desire to strengthen certain dimensions of credibility. A website perceived highly credible gets more links from others. The strength of links, in this case, the number of

[3] All TLD are registered by an authority with administrative information about the owner.

incoming hyperlinks, is an indicator of the website's credibility. Thus, website position relative to other commercial websites could be examined as a hyperlink network. Past studies analyzed the use of the number of hyperlinks between websites as an indicator of the quality of sites and found that hyperlink connectivity had a significant relationship to the expert quality judgments of sites. Also, the in-degree connectivity of a site (the number of sites that are linked to a given site) was positively correlated with judgments (Park, 2002). Further evidence can be found in more recent studies. A series of studies conducted by Persuasive Technology Laboratory at Stanford University have found that having a partner website hyperlinked may influence people perceived credibility of certain sites (Fogg et al., 2001). Thus, a website that intends to increase its credibility adds hyperlinks to credible websites. A website perceived as highly credible receives many links from others. This is what Google PageRank use as a ranking information.

2.4.3 Interpersonal and Inter-Organizational Communication

Adamic and Adar focused on university students' (Stanford University and the Massachusetts Institute of Technology) homepages and described hyperlink connections between them (Adamic & Adar, 2003). They found that some students had more than 30 incoming and/or outgoing hyperlinks while some of their schoolmates did not have any links. In order to find a connector who plays a key role in linking other homepages in the university, they measured the average shortest path between any two homepages (9.2 for the Stanford network and 6.4 for the MIT). They concluded that these results may reflect the existence of a small world network online as well as in the offline world. Besides, they examined what two students hyperlinked have in common using the content analysis of homepages.

At the inter-organizational level, Bae and Choi employed bilateral hyperlink networks among websites, to capture the structure of hyperlink communication between 402 human rights non-governmental organizations (NGOs) (Bae, S., &Choi, J.H.2000). They found that many NGOs form a hyperlink network with others according to the similar aim or activities rather than geographic location. This certainly warrants further research: How similar is the clustering of organizations based upon the content analysis of mission statements to that of hyperlink network analysis?

Hyperlink Network Analysis (HNA) derives from Social Network Analysis. We can potentially identify the social relationships by analyzing the distribution of hyperlink interconnections among websites that represent social system components such as people, private companies, public organizations, cities, or nation states (Park, n.d.2003).

2.5 Visualization

While network analysis allows for detailed examination of network characteristics, grasping its complexity requires another paradigm, in this case: visualization. Data and network visualization are part of a methodology to create powerful and insightful images with the data while carefully controlling what the user of the visualization can discover or conclude relying on its visual capability. Spatialization is defined here to transform high-dimensional dataset systems into low-dimensional spatial representations to facilitate data exploration and knowledge building (Shiffrin & Börner, 2004). Indeed, data visualization has three main goals: provide insights into a complex situation for the researcher, help verify insight validity and communicate the discoveries to a larger public (C. Chen, 2010), (DiBiase et al., 1992). Data visualization and the quite similar field of information visualization dedicated to abstract data both make use of computer capabilities to propose systematic ways to visualize data and information (Card, 1999).

Numerous charts techniques have been invented since the XVIIe (Spence et al., 2017), together with visual semiology (Monmonier, 1985), (Tufte, 1983) and implemented through visualization frameworks (Murray, 2013), (Bostock et al., 2011), (Reas & Fry, 2004) that go far beyond the traditional statistical diagrams of bart chart, pie chart and line chart.

2.5.1 Create Insights

Coming from geographic visualization is the idea that visualization for oneself is an adequate methodology to formulate hypothesis. Indeed, observing and manipulating data is an excellent way to get to know them and therefore getting insight as to how they are structured and organized. In point of fact, grasping a complex dataset from a summary of its statistical properties is not an easy task and requires a deep knowledge of said statistics. Moreover, this is not always

possible in an interdisciplinary perspective where the concepts of two or more fields have to be mastered to get a deep enough knowledge and get skills advanced enough to full comprehend the data at hand. Visualization, when carefully chosen, comes in handy to provide an overview of the data, witness structural changes when filtered and formulate hypothesis. Its mantra, overview first, filter & focus, attend to particular (Shneiderman, 1996) is a guideline for insights discovery. First one looks at the whole of the data to comprehend its extent. Then one combines data point and dimensions to test their combination and focus on local changes while looking for patterns. At last, one attends to particular to have an insight not only on the majority of data but also to the irregularities and uncertainties as they shall also be explained.

We applied this methodology with systematic visualization of every step of our methodology to lead this interdisciplinary study of diabetes related networks. In this work, visualization refers to transform high-dimensional dataset systems into low-dimensional spatial representations. That means in order to convey the information we have to scale the data so that we can display it in a two-dimensional (2D) or three-dimensional (3D) coordinate system while losing the least amount of information.

2.5.2 Network based visualizations

Our object is to reveal the network of diabetes-related websites which the most important data type is text and hypertext as well as multimedia web page contents. These data types are not easily described by numbers and therefore, most of the standard visualization techniques cannot be applied. In most cases, first a transformation of the data into description vectors is necessary. A technique for this change from textual to numerical data could be word counting for instance. A graph is a set of objects, called nodes, and connections between these objects, called edges (Keim, 2002). Any relational databases are examples for this type of datasets.

Hence, we assimilated visualizations to be map-like, exhibiting graphic elements and design characteristics of traditional maps. However, map-like visualizations are not mapping in the traditional sense, because they describe abstract information spaces, instead of geographic space. Map-like visualizations are typically restricted to two-dimensional (2D) displays, with the

possible exception of landscape visualizations (Skupin, 2002). Web networks can be represented as a 2D visualization with a sense of locality (Pfaender et al., 2006). To visually manipulate networks several options are available as visualization software evolve but only a few are capable of providing a rich visualization process such as the one describes in the previous section. Among them, one can find Tulip (Auber, 2004), Pajek (Batagelj & Mrvar, n.d.) or Gephi (Bastian et al., n.d.). After a careful review, we chose to use Gephi for its capability of producing end use map-like visualization of web network visualization plus the capacity to apply most network algorithm easily in an interdisciplinary perspective.

2.6 Semantics

One goal of this work is to extract the meaning behind the network organization to unravel diabetes web structure and its underlying social structure of actors. To extract meaning on hypertext, one can use several methodologies. We used 3 methods in conjunction to qualify the networks we work with: manual annotation, ontologies and finally Natural Language Processing.

2.6.1 Annotations / terminology

Terminology is a general word for the group of specialized words or meanings relating to a particular field, and also the study of such terms and their use. Terms are words and compound words or multi-word expressions that in specific contexts are given specific meanings — these may deviate from the meanings the same words have in other contexts and in everyday language. Terminology is a discipline that studies, among other things, the development of such terms and their interrelationships within a specialized domain (Castellví, 1999). It does this through the research and analysis of terms in context for the purpose of documenting and promoting consistent usage. Terminology can be limited to one or more languages (for example, "multilingual terminology" and "bilingual terminology"), or may have an interdisciplinary focus on the use of terms in different fields.

2.6.2 Conceptual Approach Ontologies

Ontology is the philosophical study of being. More broadly, it studies concepts that directly relate to being, in particular becoming, existence, reality, as well as the basic categories of being and their relations (Lawson, 2004). Traditionally listed as a part of the major branch of philosophy known as metaphysics, ontology often deals with questions concerning what entities exist or may be said to exist and how such entities may be grouped, related within a hierarchy, and subdivided according to similarities and differences. Medical ontology refers to the common, most general, most fundamental, highest basis, essence, or basic knowledge or theory of "about medicine" (Mjølstad, 2015).

Many ontologies have been performed in the medical field, such as neurodegenerative diseases (Cardoso & Charlet, 2017), Bilingual Ontology of Alhzeimer's disease and Related Diseases (Dramé et al., 2014), or the Nursing Coordination field, such as the NCCO's (Nursing Care Coordination Ontology) (Popejoy et al., 2015).

In the diabetes field, there already had food ontology for diabetes control, DDO (Diabetes Mellitus Diagnosis Ontology). DDO is developed within the framework of the basic formal ontology and the ontology for general medical science to represent entities in the domain of diabetes, and it follows the design principles recommended by the Open Biomedical Ontology Foundry[4]. Currently, DDO contains 6444 concepts, 48 properties, 13,551 annotations, and 27,127 axioms. DDO can serve as a diabetes knowledge base and supports automatic reasoning. It represents a major step toward the development of a new generation of patient-centric decision support tools (El-Sappagh & Ali, 2016). However, there is still none ontology addressed the stakeholders or the social networks in the diabetes context. This manuscript is mainly focus on the visualization of the networking on diabetes-related websites, hence, we need to find another way to tag the websites semantically.

4 http://www.obofoundry.org/

HeTOP[5], the Health Terminology/Ontology Portal, is a tool dedicated to both human beings and computer to access and browse biomedical terminologies or ontologies (T/O). HeTOP is a cross-lingual terminology server which is already been used by 500 unique machines per day, mainly by librarians, translators, students and physicians (Julien et al., 2013). The translations of terms and the interoperability between T/O such HPO are also a major leverage for the quality of the data and terminologists and ontologists could find in HeTOP a great opportunity to deal with the lexicons quality.

2.6.3 Natural Language Processing

Natural Language Processing (NLP) (Manning et al., 1999) is a set of technics to manipulate language representation automatically used in many fields (Al. (eds et al., 2004), (Ingrid E. Fisher et al., 2016). This is the case of hypertext, which is itself made of text assimilated to a written language. We can apply NLP to large hyper textual resources to automatically extract their vocabulary, key concepts and topics. The field of NLP is vast and this thesis purpose is not to explore all its capabilities but to focus on a practical case where manual annotation is a labor-intensive process that can benefit from an automatic text analysis process. For such an analysis, our goal was to focus of topic modeling (Uys et al., 2008) to supply annotation with valid topics extracted from a corpus of text, that a human could review.

To perform a topic modeling from a set of documents in a corpus, several steps have to be followed. First the text has to be cleaned to remove unnecessary variation of the same word and only keep the canonical version known as a lemma. Similarly, unwanted most frequent word (stop words) shall also be removed to focus on meaningful words only. To clean efficiently, a process known as lemmatization, we uses a Part-Of-Speech (POS) decomposition of every sentence that most of NLP tools dispose of (Al Omran & Treude, 2017).

Once every document is cut into words and lemmatized, we transform the collection of documents (a corpus) into a matrix of token counts. Some methods such as TFIDF (Ramos, n.d.) put a threshold to remove the most frequent and least frequent words but multilingual

[5] https://www.hetop.eu/hetop/

interdisciplinary web corpora have a very large distribution of word frequency and thresholding is usually not necessary if not counterproductive. The matrix of words count is then decomposed similarly as a dimension reduction. But in this case the decomposition yields a set of features (group of words) that describe the corpus. Several algorithms of matrix decomposition exist such as Nonnegative Matrix Factorization (NMF) (Cichocki & Phan, 2009), or the most common for topic modeling is the Latent Dirichlet Allocation (LDA)(Hoffman et al., 2010).

With this state of the art in mind, we started the web exploration starting with the harvesting of diabetes web data.

3 Chapter 3 Mapping the Hyperlink Structure of Diabetes-Related Websites

In this Chapter, we present the methodology to harvest and visualize the map of hyperlinks structure on diabetes communities. At the beginning, we present the two state-of-arts tools we used in our work, Hyphe and Gephi. Then, we go step by step to explain our procedure and how we leverage these two tools to extract the final map with 430 diabetes-related websites and 6587 hyperlinks.

3.1 Data collection: Web Crawler

The key factors of the success of the World Wide Web are its large size and the lack of a centralized control over its contents. These characteristics are also the most important source of problems for locating information. The Web is a context in which traditional Information Retrieval (IR) methods are challenged and given the volume of the Web and its speed of change, the coverage of modern search engines is relatively small. Moreover, the distribution of quality is very skewed, and interesting pages for a given user are scarce in comparison with the rest of the content (Castillo, 2005).

Web crawling is the process used by search engines to collect pages from the Web (Baeza-yates & Castillo, 2002). A web crawler, also known as a robot or a spider, is a system for downloading web pages in bulk. It is an automated program or script that scans or “crawls” web pages methodically to create an index of the data which is set to look for. This process is called Web crawling.

Web crawlers are used for different purposes, but essentially a web crawler is used to collect/mine data from the Internet. Most search engines use it as a means of providing up-to-date data and to find what is new on the Internet. Analytic companies and market researcher use web crawlers to determine customer and market trends in a given geography.

In this work, we need a crawler to harvest websites and webpages, capable of handling tagging and appropriate for somebody not capable of programming. We found one tool named

Hyphe can meet all these requirements and it was developed for research with interdisciplinary researchers in mind. Hence, we use Hyphe which is a web crawler developed by the Médialab department of Sciences Po. Sciences-Po is an international research university located in Paris, ranking among the finest institutions in the fields of humanities and social sciences. Médialab is the tenth research center of Sciences-Po to help social sciences and humanities take full advantage of the huge amount of data made available by digitization.

As Hyphe is a topic-specific web crawler, made for sociologist who tries to understand online communities, it is an ideal tool for us to collect data of diabetes-related websites to understand diabetes online communities. In particular, Hyphe replaced the notion of website with the more flexible notion of "web entity" introduced earlier to better adapt to web heterogenous structure. Here we mainly focus on how Hyphe can be used to crawl the diabetes-related websites instead of explaining how Hyphe works in deep to crawl Internet.

3.1.1 Description of Hyphe

Hyphe is a tool to build web corpus by crawling web pages and generating networks between "web entities", connected with each other using hyperlinks (Ghemawat, n.d.). Web entity is referred to an entity of homogenous content often assimilated to a website. In Hyphe, users choose how web pages are grouped. Web pages can be grouped by URL when they share the same domain name and Top Level Domain (TLD) as follow: domain.tld, or the same subdomain.domain.tld etc. For example, we can group twitter.com, twitter.com/BeyondType1, twitter.com/JDRF, twitter.com/search, twitter.com/share into one web entity defined as twitter.com or we can check the boundaries of each web entity before creating it. (See figure 8) Similarly, we can merge different facets of people's presence on the web: personal blog, Twitter and Facebook accounts into a single web entity. This concept is more flexible and accurate than the intuitive but vague notion of "website". Website is the web entity by default on Hyphe but this can be tuned to adapt to the web heterogeneity of content structure. Thanks to the concept of web entities, the web pages are better organized and easily to be recognized.

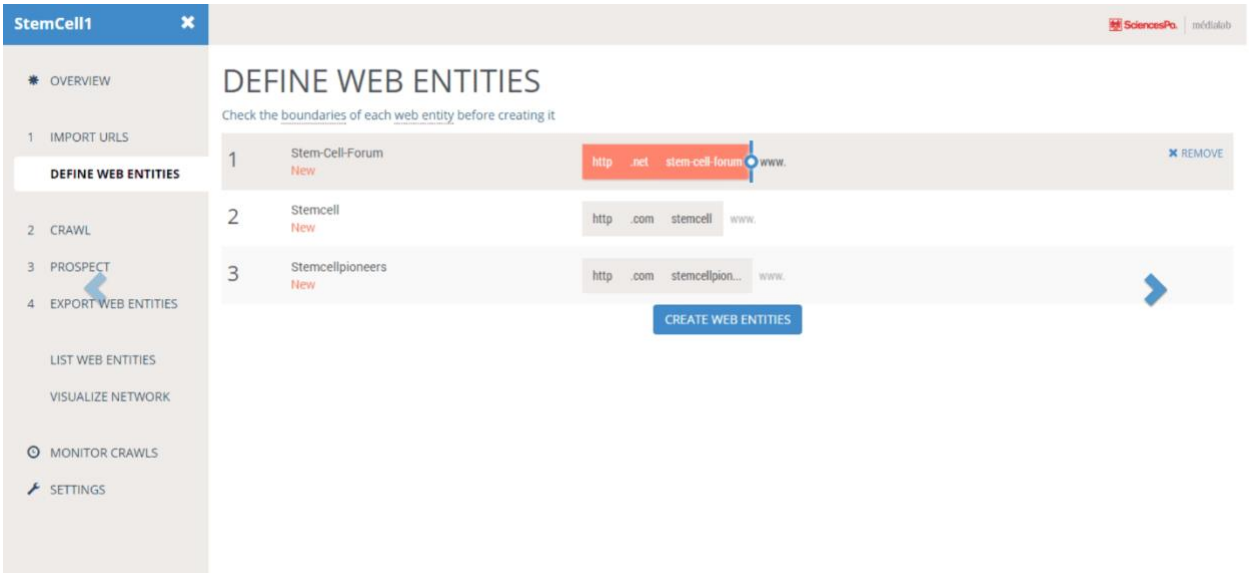


Figure 8 User interface of web entities definition in Hyphe.

Hyphe is user-centric. It does nothing unless commanded by the user, and provides a lot of feedbacks for monitoring its processing. (see figure 9) Hyphe supports a step-by-step expansion method where web entities are curated before being crawled so that hyper connected websites are not added to the corpus unless necessary to the research, thus keeping topic drift in check. The curation is done through a system of exclusive web entity statuses:

- DISCOVERED when the web entity is detected by the crawler itself, however the user has not taken any decision.
- IN when the user has explicitly accepted it.
- OUT when the user has explicitly rejected it: Websites providing many links to tools or software or popular social network that are not related to the topic at hand but rather offer another medium opportunity (ex linkedin.com, Instagram.com, google.com, etc.)
- UNDECIDED when the user has suspended his decision, often because corpus selection criteria apply ambiguously to it.

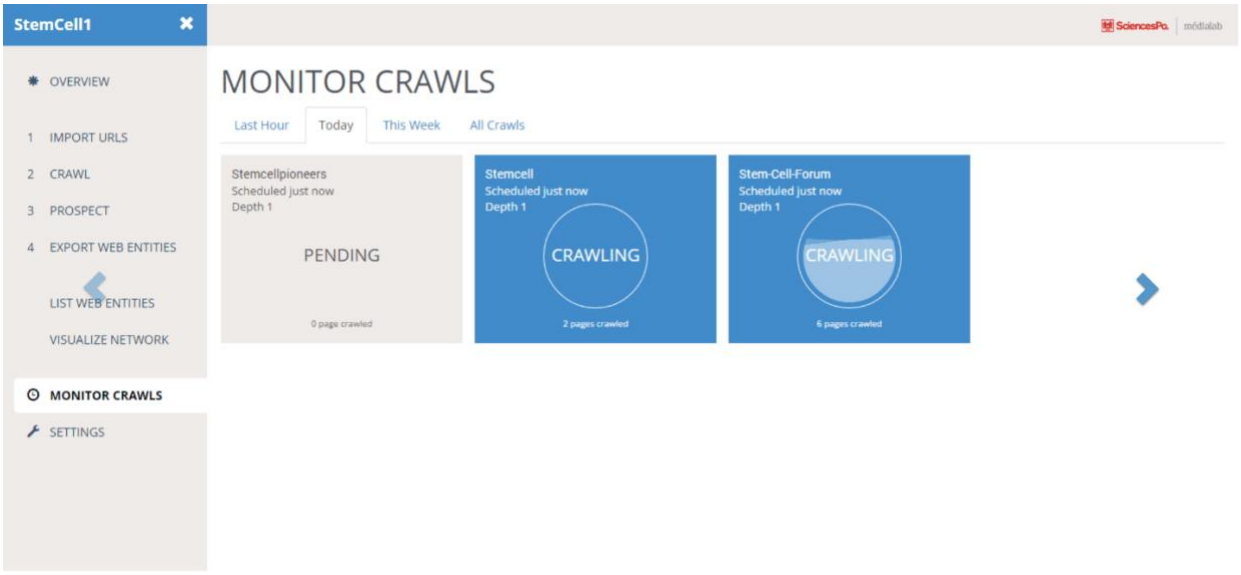


Figure 9 User Interface of monitoring crawling in Hyphe.

To optimize the curation, the interface dedicated to corpus expansion displays the list of DISCOVERED entities from the most to the less cited. (see figure 10) The user can then reject (set to OUT) the hyper connected websites while accepting (setting to IN) and crawling the relevant ones. This approach allows the user to reject pages with higher degrees of connection and resist the attraction of the hyper connected nodes and ultimately preventing the snowball effect. According to Cambridge Dictionary, snowball effect means a situation in which something increases in size or importance at a faster and faster rate (Ghemawat, n.d.).

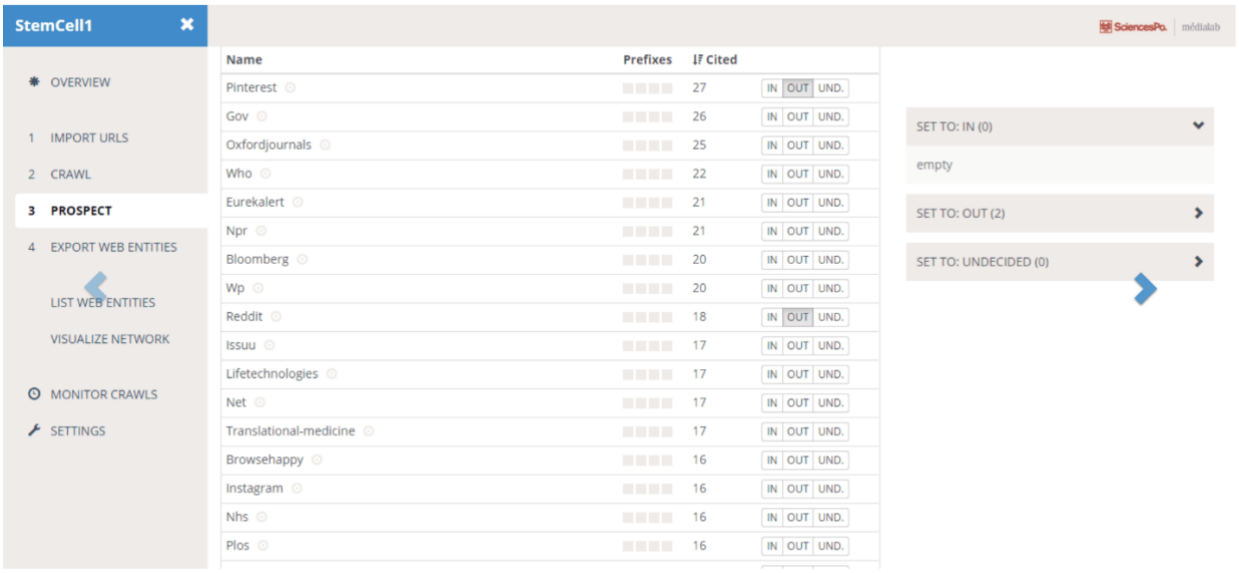


Figure 10 User Interface of the web entity statuses in Hyphe.

Hyphe enables the user to estimate the exhaustiveness of the corpus by exploiting the hyperlink structure of the web. The step-by-step expansion method leads to a systematic analysis of the most cited DISCOVERED web entities, up to a certain threshold. Provided that every accepted entity has been properly crawled, that threshold conveys the approximation of exhaustiveness that has been performed in the corpus. The lower the threshold (relatively to the in-degree distribution), the better the approximation of exhaustiveness. It helps users manage the tradeoff between corpus quality and time spent on prospection. In the end, Hyphe can offer the networks roughly. (see figure11) So far, Hyphe is not adapted to gathering more than a few million pages and 100,000 web entities.

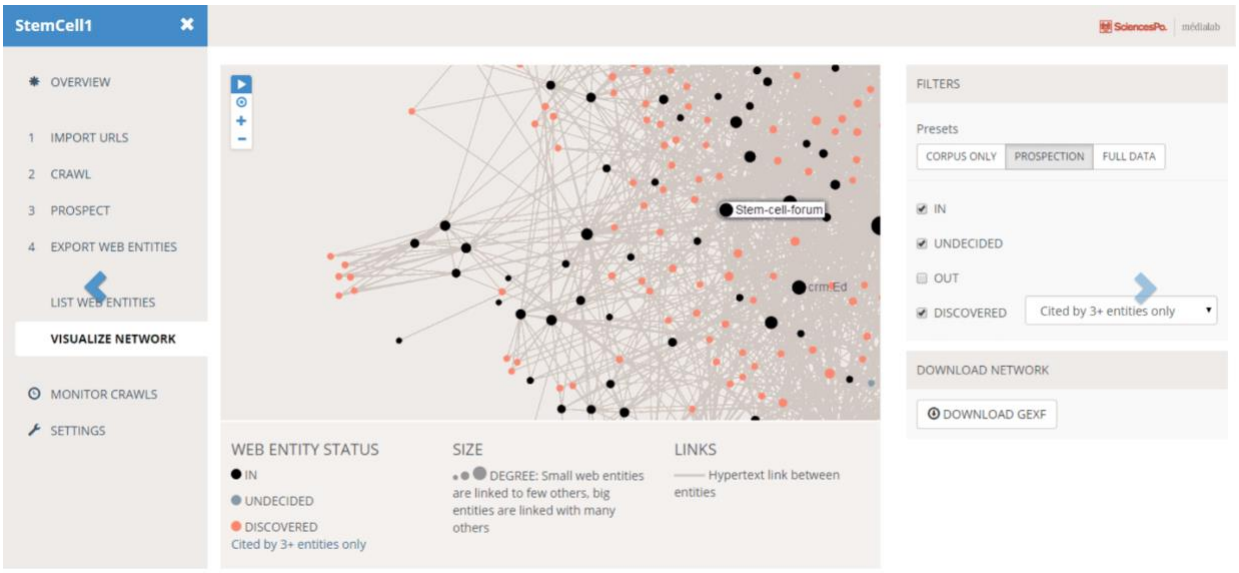


Figure 11 User interface of the visualization network in Hyphe.

3.1.2 Use cases of Hyphe

Hyphe has been used and the results are published in social science. Hyphe allows to methodically gather web entities and visualize the network aiming at data journalists (Venturini et al., 2017). Since journalists have so far made little use of the analytical resources provided by networks, the researcher studied how to conduct "visual network exploration" by Hyphe in the context of data journalism in order to explore, describe and understand large and complex relational datasets.

Tommaso Venturini and Marthieu Jacomy took a recent example from journalism, namely a catalogue of French information sources compiled by Le Monde's The Decodex as a starting point to feed Hyphe. Then they examined how "visual network exploration" can address the problem that the recent spread of digital media has increasingly confronted journalists with information coming not only in the traditional form of statistic tables, but also of relational databases. The good visual exploration of networks is an iterative process where practices to demarcate categories and territories are entangled and mutually constitutive (Venturini et al., 2017). In that work, they borrow the more familiar vocabulary of geographical maps to interpret and characterize graph structure and properties by position, size and hue. They showed on the knowledge-making capacities of Hyphe and how these compare to the insights and instruments

that journalists have used in the Decodex project.

In addition, Tournay identified and studied the community of stem-cells on the web, observing how the link structure impacts the "propagation of meaning between sociologically disparate actors" (Tournay, 2016). A team of librarians and sociologists delineated the debate on climate change on the web, notably measuring the unexpected large presence of climate skeptics (Jacomy et al., n.d.). Even more, people tried to study through the "Alternative for Germany" (AFD) Facebook wall which has already become one of the largest right-wing forums on the German-speaking internet to see how social media plays a crucial role for the party's mobilization strategy (K. Arzheimer & CC Berning, 2015). Yet, Hyphe has never been used in public health care to assess chronical disease online communities' structure. Moreover, even if Hyphe does include basic network visualization capabilities, it quickly reaches its limits as the networks grows beyond one hundred nodes.

3.2 Graph Visualization

In the aim of understanding networks, graph visualization is an effective approach to spatialize the relationships (Sporns et al., 2004). Using a visual display helps identify features in a network structure and data while relying on human very efficient perceptual abilities.

Graph Visualization can only be fully understood when visualization becomes an integral part of user activity. This includes such obvious applications as counter-terrorism work or the development of improved Web search engine interfaces. Telecommunications companies attempt to find patterns in millions of phone calls through visualization. Private industry also hopes to use visualization to detect emerging technological trends from research literature in order to gain a competitive advantage. Funding agencies would like to determine which research grant applications show the most promises.

In recent years, there have been a growing number of events dedicated to the type of research within which visualization is prominently featured, organized by the National Academy of Sciences the National Institutes of Health, the National Security Agency and other public and private entities (Shiffrin & Börner, 2004).

We can identify some main requirements for a network exploration tool: high quality layout algorithms, data filtering, clustering, statistics and annotation. In practice, these requirements must be included in a flexible, scalable and user-friendly software. Since the typical output of Hyphe is a network of web entities to be analyzed through network analysis software such as Gephi, we choose to use Gephi as graph visualization tool to present the final map of diabetes-related websites. Focusing on analysis clarity and on modern user interface, the Gephi brings the high-quality network visualization to both experts and inexperienced audience (Bastian et al., n.d.).

Gephi is an open source network exploration and manipulation software. It can import, visualize, spatialize, filter, manipulate and export all types of networks. One of its main functions is the ability to display the spatialization process, which is designed to convert the network (an abstract raw structure) into a meaningful map. It does it by integrating three essential aspects – network spatialization, network statistics and visualization design – into a single dedicated interactive WYSIWYG (what you see is what you get) interface that do require computer programming experience compared to its counterparts (Pavlopoulos et al., 2017) (Telea, 2014). Gephi is capable of spatializing large network (i.e. over 100,000 nodes) and allow users to manipulate it to get visual insight. Users can modify vertices, edges, vertices labels and edges labels by adjusting their size and color with manually input values or by associating size and colors to predetermine properties or statistics. Network visualization itself relies heavily on the layout algorithms that determine the spatial position of vertices in a two-dimensional or three-dimensional space (Purchase et al., 1996).

One of the most prominent layout of dense network is force vector layout as is tries to distribute vertices in space while having edges of a similar size. This allows users who take visual decisions based on the layout to not overly interpret edges length which in this case represent hyperlinks. Hyperlinks themselves are not ranked in importance and an equal edges length algorithm focused on vertices readability instead is quite adequate for our web network analysis task. Force Vector uses edge as spring-like forces and nodes as repulsive force to simulate the system and bring it to a minimum energy equilibrium (Kobourov, 2012). ForceAtlas2 is gephi default layout algorithm. It is a force-oriented layout that was designed with web networks in mind

(Jacomy et al., 2014) and take into account the scale free property of these networks to obtain a balanced layout where scale free network with traditional force vector layout would otherwise show a very hard to read dense core and disparate periphery. The text module can show labels on the visualization window from any data attribute associated to nodes.

The user interface (see Figure 12) is structured into Workspaces, where separate work can be done. Great attention has been taken to the extensibility of the software. An algorithm, filter or tool can be easily added to the program with little programming experience to match domain specific needs, such as topological network or web network. Sets of nodes or edges can be obtained manually or by using the filter system. Filters can select nodes or edges with thresholds, range and other properties. In practice filter boxes are chained, each box takes in input the output of the upper box. Thus, it is easy to divide a bi-partite network or to get the nodes that have an in-degree superior to 5 and the property “type” set to ”1”. Because the usefulness of a network analysis often comes from the data associated to nodes/edges, ordering and clustering can be processed according to these values (Bastian et al., n.d.).

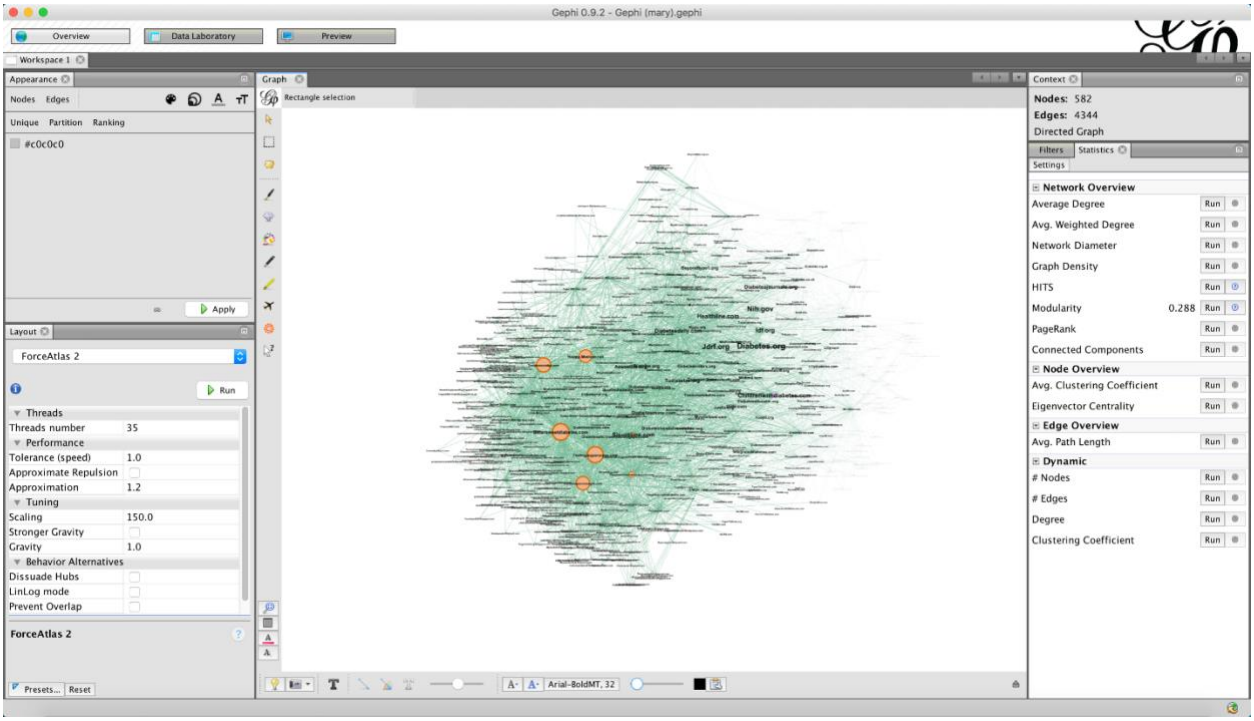


Figure 12 An interface of Gephi beta version 0.9.2 with an ongoing analysis bringing a specific focus (orange nodes) on a set of targeted websites having a peculiar set of tags (see chapter 6 for similar tag related network analysis).

Though networks can be explored in an interactive way with the visualization module, it can also be exported as a PDF file or a Simple Vector Graphic (SVG) for expert designer willing to customized the map beyond Gephi capabilities in a specific software such as illustrator™. A powerful SVG exporter named Rich SVG Export is included in Gephi. Many options are offered to users to set the design of nodes, edges and labels. Techniques are developed to increase networks clarity and readability.

In this context, we have developed a methodology to uncover the online organization of diabetes websites. We leveraged and made best out of these tools to collect information using Hyphe and then visualize their properties and links using Gephi to produce the network on diabetes websites and finally provide key insights about this community.

3.3 Materials and methods

The proposed methodology consists in a workflow embedding the following milestones:

- collecting diabetes community websites
- analyzing the resulting websites structure
- visualizing the diabetes community network as a navigational map

These milestones can be decomposed into a more detailed process potentially applicable to any online community analysis (see figure 13). The methodology thus follows the following steps:

- gathering a list of starting websites
- creating a pre-crawl corpus
- consolidating the pre-crawl corpus
- launching and performing a full-scale crawling
- extracting the websites network
- visualizing the network structure

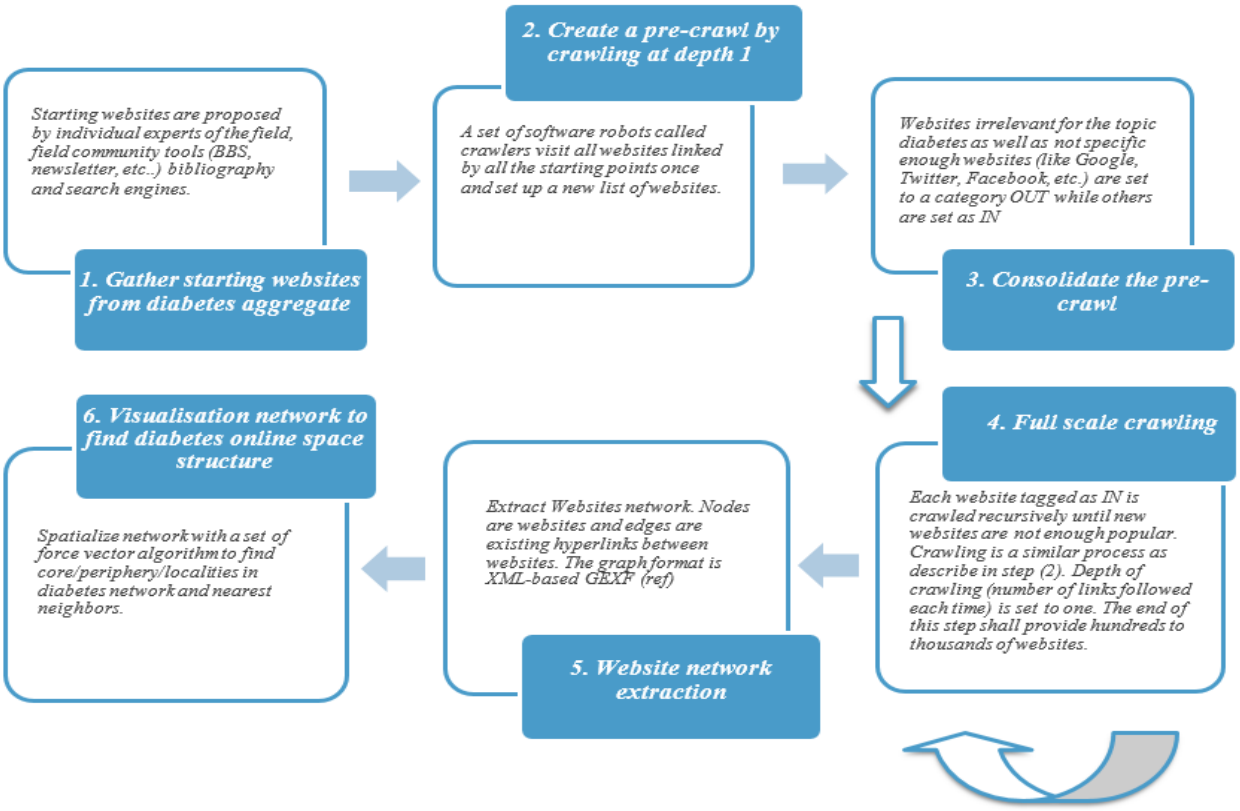


Figure 13 Illustration of the workflow.

These six steps are based on the following decision criteria.

Collection of starting websites: In order to explore an online community consisting in topic related websites, we need to define entry points into the community. These entry points will serve as a gate into the community by following the points links with a robot crawler carefully controlled. As the objective is to explore all aspects of the community, the entry points shall reflect different points of view, opinions or topics the community might address so as to make sure to capture enough gates leading to the potential different parts of the final network. To ensure the diversity of entry points we selected a variety of sources starting with (a) two or more experts of the topic suggesting websites (for the diabetes topic, these experts were found in people who are living with diabetes and are also working in diabetes area for several years); (b) using various queries about diabetes on Google to ensure different results and (c) using diabetes community social network pages suggestions.

Choice of depth and the criteria to stop crawling: Once the starting web entities are set, Hyphe offers the possibility to follow their hyperlinks (html markup <a> for anchor) automatically by detecting them in web pages. Hyphe have three depths for crawling, from 1 hyperlink away from the original web entity to 3 hyperlinks away to (3 clicks) depending on how deep the web entities shall be explored. (see figure 14) As the web network follows a power law degree distribution, a crawl at depth 3 can potentially gather thousands of websites making it overwhelming for a human to properly assess each and every one of their individual membership to the targeted diabetes community. In this thesis work, to avoid too much noise, we decided to consider only a slower depth level of 1, repeated several times if needed but only for targeted websites whose contribution to the topic is substantial. Moreover, the criteria to stop crawling new web entities depends on how much websites related to diabetes are showing up in the final results. If no new diabetes related entity shows up after the crawling, then we considered to have reach a local frontier in the diabetes community and proceed with others web entities until no more entities are left to be un-crawled. Then all topic frontiers are discovered. At last, considering the huge numbers of web entities, we did not consider in this first study the entities mentioned only once. It is a web network property that noticeable entities should be linked to by several neighbors. A web entity mentioned only once have little chance to be of importance (Brin & Page, 1998). We considered a minimum of 3 hyperlinks to an entity which we call it “popular” website for it to be eligible for an additional depth 1 crawling.

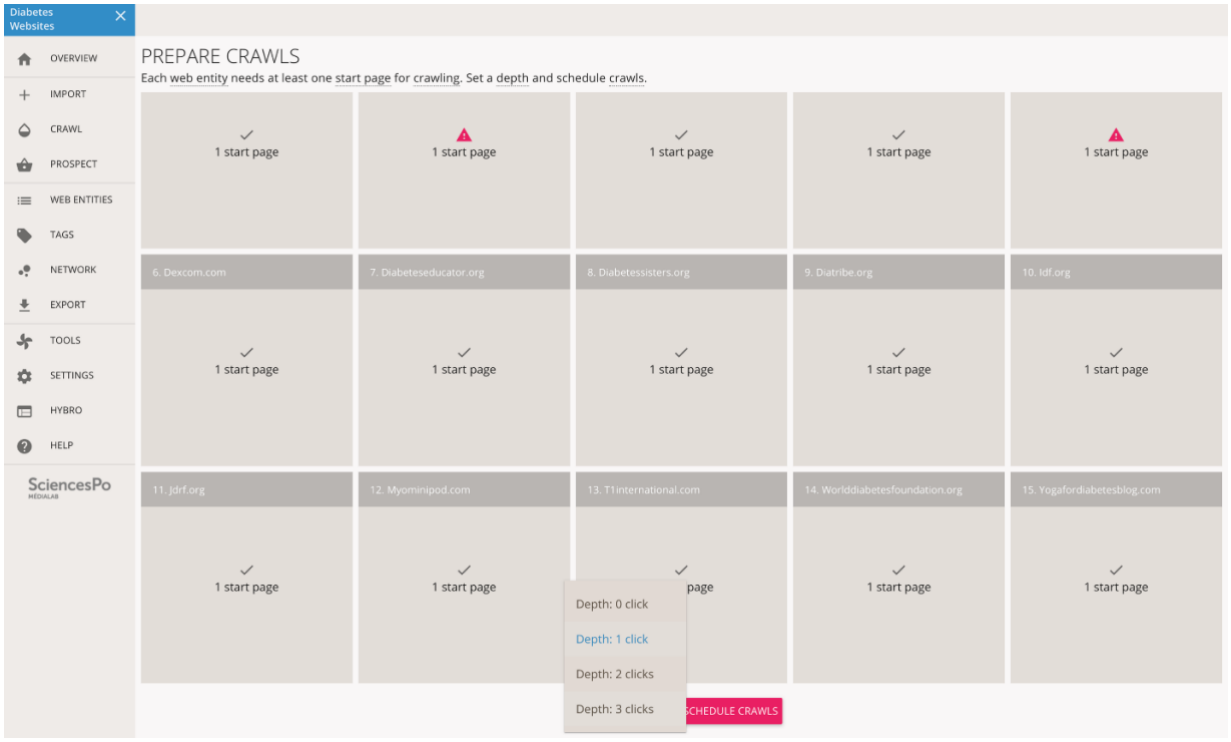


Figure 14 Preparation crawls from depth 0 to depth 3 with starting webpages and the pull-down menu of the depth options.

Cleaning process to feed each iteration of step 4 with an IN database: To clean the database after each one depth level iteration, we used two ways to make a decision and classify the websites as: IN, OUT or UNDECIDED. In this work, we classified the contents of the web entities strictly related to diabetes as IN; when nothing indicates a potential diabetes related content, we classified it as OUT; finally, ambiguous websites that mention diabetes among other topics were categorized as UNDECIDED waiting for further in-depth analysis to lift the veil on their classification. The first way to classify the entities is to automatically select in the database the URL string containing the word diabetes and default them as IN. The second way to classify the entities is to select manually the class after a review of the content of each website. This is done by opening its URL in a browser. After the classification, we can filter the database to only retain the IN websites. As shown by figure 15, Youtube.com, Google.com, Instagram.com, Amazon.com are not specific enough websites for diabetes, we all set them to a category OUT while others are set as IN.

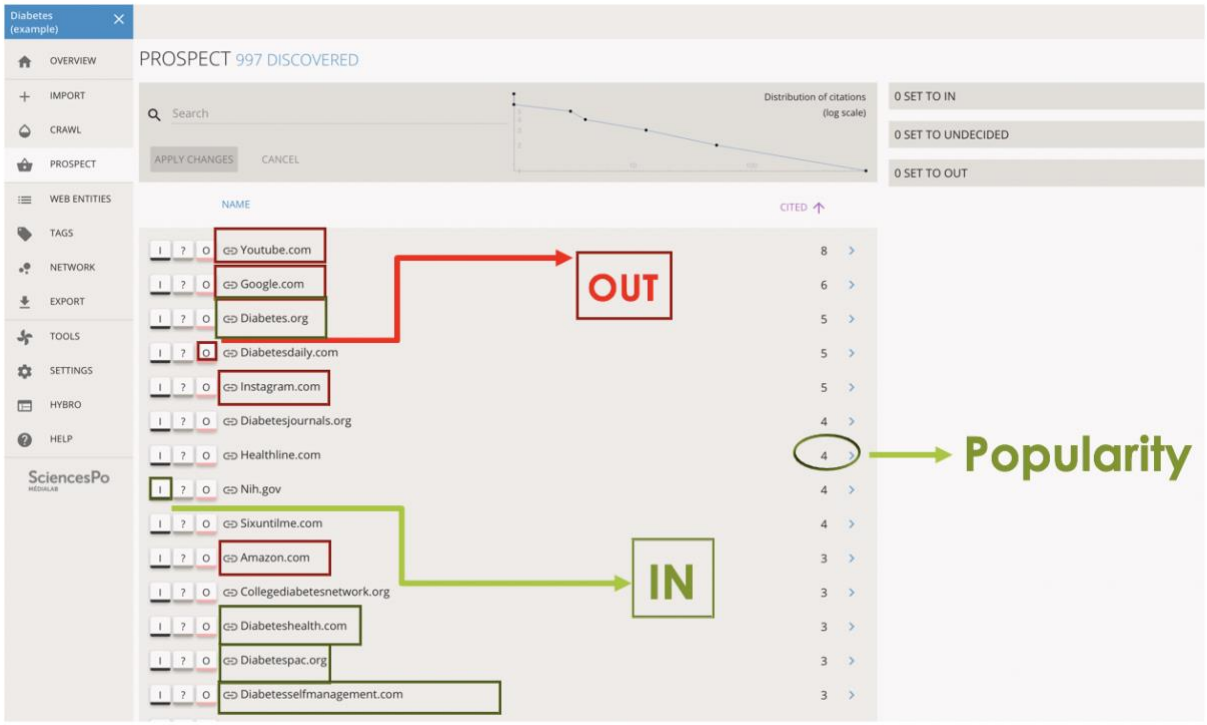


Figure 15 Database clean process to retain the IN websites, "I" delegates IN, "O" delegates OUT and the number delegates the popularity of the website.

Choice of the final pattern in step 6 (Gephi): Once the "pre-crawl" was created by crawling the starting websites at depth 1 and cleaning them, we applied the crawl & clean process again several times to get the full-scale crawling (step 4). (see figure 16) Each web entity regarded as IN is crawled recursively until no new websites are popular enough which are citied above 3 times to be included or are all off-topic (OUT or UNDECIDED). The end of this step provides with hundreds of IN websites and their hyperlinks. Then we exported the websites network as a graph in the XML-based GEXF format. In the final network, nodes are websites and edges are existing hyperlinks found between websites (step 5) weighted by count of occurrences. Lastly, we imported the graph to Gephi to visualize the hyperlink structure of diabetes online communities (step 6). In order to do so, after importing the network, we used a force directed "Force Atlas 2" layout well adapted to scale free networks to spatialize the nodes and their edges. The edges represent hyperlinks only. Therefore, the proximity of the nodes is the result of the force-vector based layout algorithm only and do not demonstrate any physical proximity whatsoever.

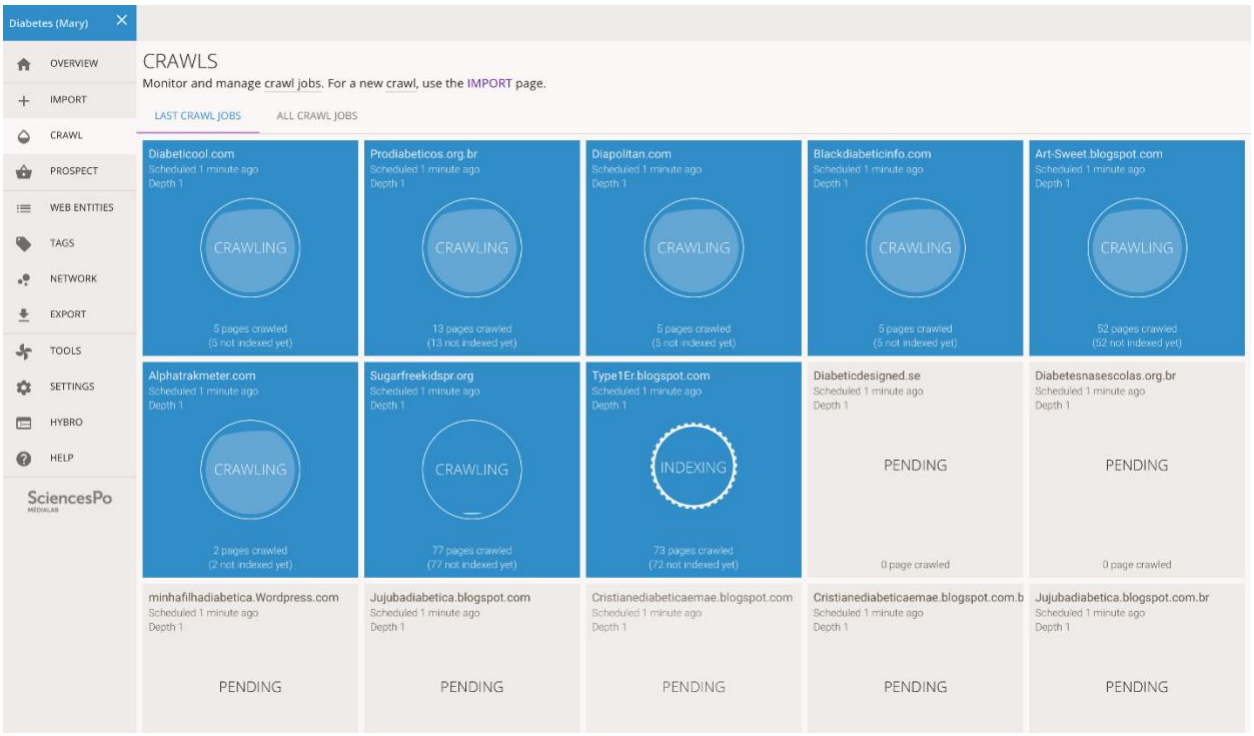


Figure 16 Interface of Full-scale crawling by Hyphe, "crawling" means the websites are under crawling process, "indexing" means ongoing and "pending" means waiting for being crawled.

3.4 Results

As the starting websites were proposed by individual experts of the field, by field community tools (BBS, newsletters, etc.) bibliography and general search engines like Google, we combined all the resources and picked up most well-known 15 websites which represent different aspects in the diabetes world as the starting points to feed Hyphe. The table 1 describes these 15 websites.

Table 1 Starting websites for crawling.

Different Aspects	**Websites**
1 portal website	http://www.childrenwithdiabetes.com
1 well-known patient's blogger	https://yogafordiabetesblog.com
1 local association focus on diabetes educator	https://www.diabeteseducator.org
1 study focus on diabetes globally beyond Novo nordisk	https://www.dawnstudy.com
2 diabetes publications	https://asweetlife.org https://diatribe.org
2 leading pharmaceutical companies	https://www.myomnipod.com https://www.dexcom.com
2 international associations	https://www.jdrf.org https://www.idf.org
5 charities	https://www.t1international.com https://beyondtype1.org https://www.worlddiabetesfoundation.org https://www.diabetessisters.org https://diabetesdestiny.com

After the web crawler Hyphe visited all websites linked by all the starting points once, we set up a new list of websites and use the cleaning process to make sure we only keep the IN

websites which are strictly talking about diabetes. Once Hyphe discovered as many web entities as possible and their hyperlinks, we get the final number of the DISCOVERED web pages in Hyphe overview part. Each of the discovered website then is classified as IN/OUT/UNDECIDED. In the end, we successfully captured 430 IN websites with 6587 hyperlinks, each website contains information about diabetes or some aspects of it together with some basic tags on the nature and content of these websites. Figure 17 is the diabetes hyperlink structure as visualized by Hyphe directly. This graph is "messy" and while it shows a strong core component, there is no visible underlying structure so far.

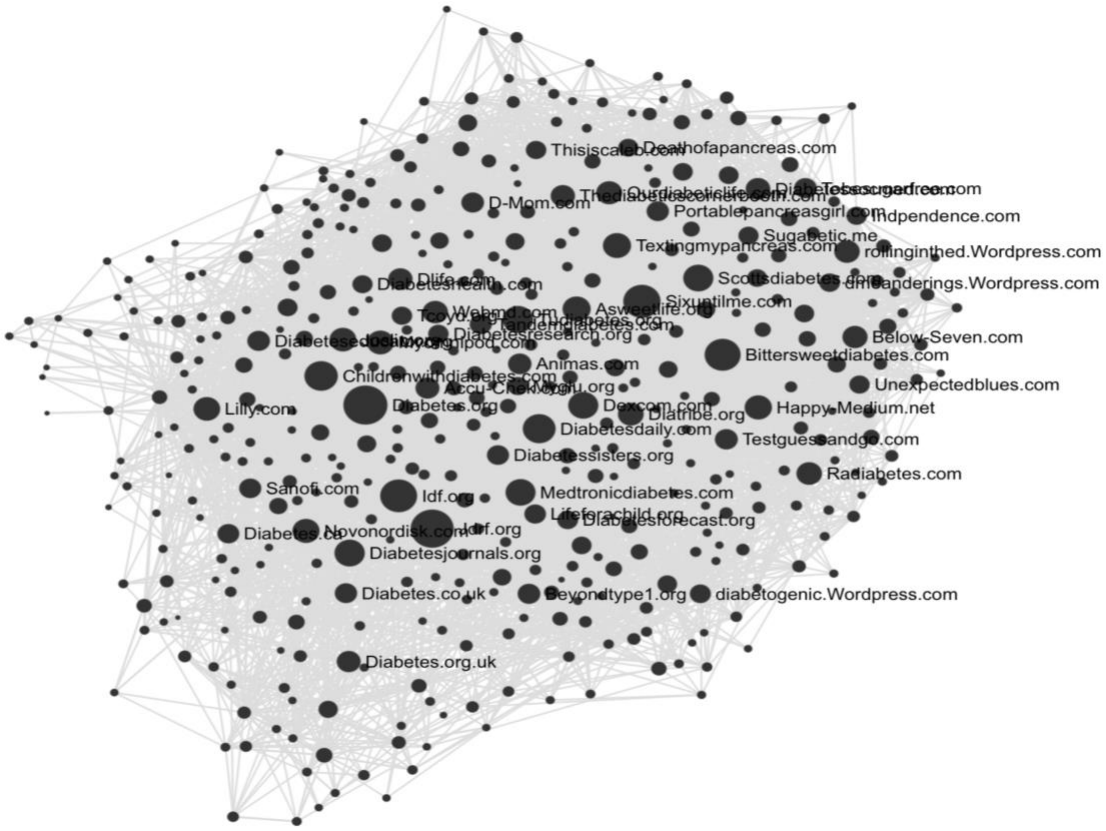


Figure 17 Diabetes hyperlink structure with basic force directed layout.

As already mentioned in Chapter 2, the World Wide Web represents a very large network for which topological information are accessible. The nodes of the network are the documents (webpages) and the edges are the hyperlinks (URLs) that point from one document to another. The interest in the web as a network has boomed after it has been discovered that the degreed

distribution of the webpages closely follows a power-law distribution, even if only extract of the web are observed (Al- bert, Jeong, Baraba´si 1999, Kumaret al. 1999).

Since the edges of the web are directed, the network is characterized by two degree distributions: the distribution of outgoing edges, $P_{out}(k)$, signifies the probability that a document has k outgoing hyperlinks and the distribution of incoming edges, $P_{in}(k)$, is the probability that k hyperlinks point to a certain document. Several studies have established that both $P_{out}(k)$ and $P_{in}(k)$ have power-law tails:

$$P_{out}(k) \sim k^{-\gamma out} \text{ and } P_{in}(k) \sim k^{-\gamma in} .$$

Albert, Jeong and Barabasi (1999) have studied a subset of the web containing 325,729 nodes and have found $\gamma_{out} = 2.45$ and $\gamma_{in} = 2.1$. Kumar et al. (1999) used a 40 million document crawl by Alexa Inc., obtaining $\gamma_{out} = 2.38$ and $\gamma_{in} = 2.1$ (see also Kleinberg et al. 1999). A later survey of the web topology by Broder et al. (2000) used two 1999 Altavista crawls containing in total 200 million documents, obtaining $\gamma_{out} = 2.72$ and $\gamma_{in} = 2.1$ even a considerably larger scale (more than 600 times larger). Adamic and Huberman (2000) used a somewhat different representation of the web in which each node represents a separate domain name and in which two nodes are connected if any of the pages in one domain linked to any page in the other. While this method lumps together often thousands of pages that are on the same domain, representing a nontrivial aggregation of the nodes, the distribution of incoming edges still followed a power-law distribution with a comparable coefficient $\gamma^{dom} = 1.94$.

In our work, diabetes-related websites network containing 430 nodes and 6587 hyperlinks, we found that the distribution of outgoing and incoming links both followed a power-law distribution with coefficients $\gamma_{out} = 3.78$ and $\gamma_{in} = 1.99$ (see figure 18). Both coefficients are in range with typical values found in other web related networks. This confirms that the network we retrieved through Hyphe is well in range with its counterparts. Furthermore, it confirms that Hyphe is supposed to and yield a proper web network.

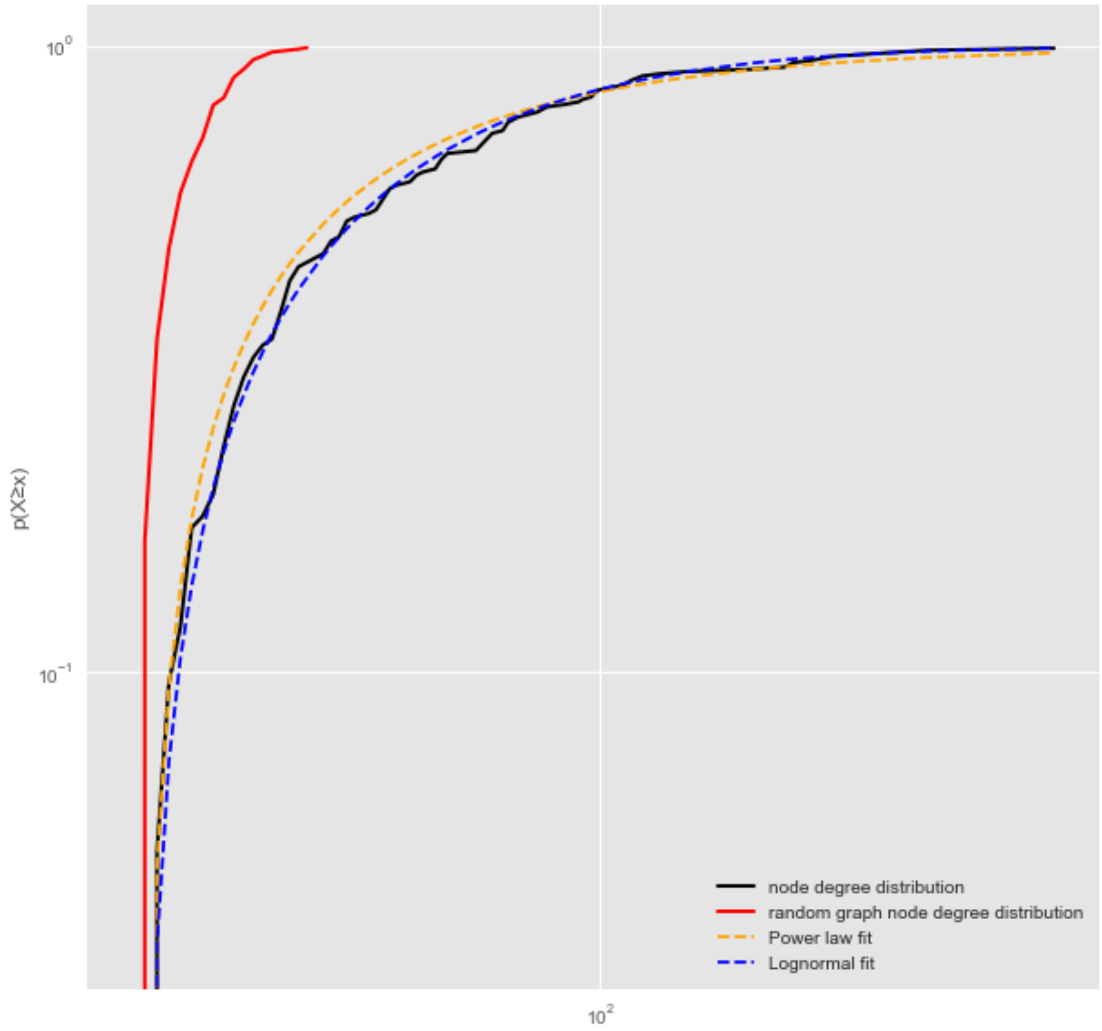


Figure 18 430 diabetes-related websites nodes degree distribution.

Despite a very large number of nodes, the web still displays a small world phenomenon property. This was first reported by Albert, Jeong and Barabasi (1999), who found that the average path length for a sample of 325,729 nodes was 11.2 and predicted, using finite size scaling, that for the full web of 800 million nodes at that time that would be around 19. Subsequent measurements of Broder et al. (2000) found that the average path length between nodes in a 200 million sample of the web is 16, in agreement with the finite size prediction for a sample of this size. Finally, a domain level network displays an average path length of 3.1(Adamic 1999). In our study, the average path length is 2.4. This is a relatively small average path that is partly corollary of our step by step crawling. But it also proves that the topic specific diabetes web network possesses a small work phenomenon and is organized in a structure where actors are close to one

another. This would not be the case with a bipartisan topic where the path length would be much larger as in the web in general even if it stays in a range lesser than 10. Therefore, it makes sense to study the web of diabetes as the digital traces of an organized social group.

The directed nature of the web does not allow us to measure the clustering coefficient using scaling relationships (Enquist et al., n.d.). One way to avoid this difficulty is to make the network undirected, making each edge bidirectional. This was the path followed by Adamic (1999) who studied the web at the domain level using a 1997 Alexa crawl of 50 million webpages distributed between 259,794 sites. Adamic removed the nodes which have only one edge, focusing on a network of 153,127 sites. While these modifications are expected to increase somewhat the clustering coefficient, she found $C = 0.1078$, orders of magnitude higher than $C_{rand} = 0.00023$ corresponding to a random graph of the same size and average degree. In our work, we found that average clustering $C = 0.3931$, orders of magnitude is higher than average clustering coefficient of a random graph $C_{rand} = 0.0705$. It shows that nodes in our web network have a tendency to form clusters and group together much higher than an equivalent random network. Therefore, the non-natural shape of our network has to be the result of a structuration under impulsion of its actors.

These figures prove our hypothesis that diabetes online communities have their own space in the digital world. All stakeholders involved in diabetes have connections of some sort that can be revealed in the World Wide Web as an organized world of communities rather than a randomly organized network. At the same time, we also can prove Hyphe is an appropriate tool to offer us a web network since we do extract the websites to feed in Hyphe.

In order to visualize the diabetes online communities in more details, we created the visualization in Gephi. Then to provide a first view at the sub-communities inside the diabetes community we applied a community detection algorithm aimed to detect clusters of similarly connected nodes (Blondel et al., 2008). It does so by maximizing the quality metric known as modularity over all possible partitions of a network (Newman, 2006).

Modularity measures the difference between the edge density in the partition and a randomized graph with the same number of nodes and the same degree distribution (Chen et al., 2014). Then a color map is associated to 5 discovered communities class resulting in the figure 19.

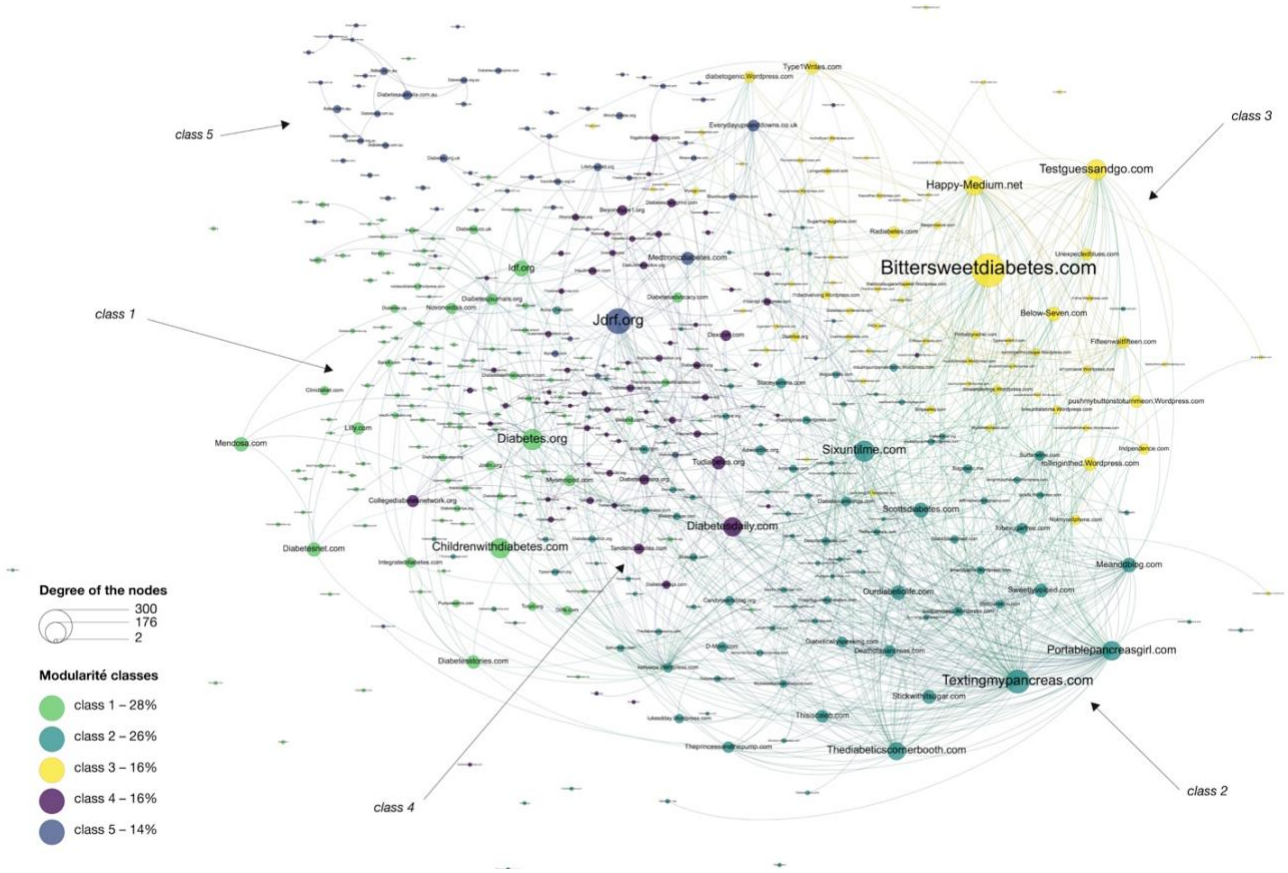


Figure 19 Diabetes hyperlink structure with force atlas 2 layout and communities.

The result shows that diabetes online community has its own space and is organized with clusters and distribution of websites of relative importance. More analysis is necessary to show exactly what is the topic (see Chapter 4) but we get a general idea that they don't belong together by chance. It is remarkable that the algorithm working only on node probability highlights these communities that in real life share a common interest when we read their content. If we want to have some general ideas about which community mainly talking about which topic, we need to manually or semantically annotate each website using terminology of the diabetes fields would be a good way towards explanation the communities' contents. Moreover, we used a degree centrality

algorithm (PJ Carrington, 2005) which assigns an importance score based on the number of links held by each node to get several central websites in our 430 websites diabetes digital. Node Degree reveals the number of IN and OUT hyperlinks which are from 2 to 303 related to one web entity. Among them:

- bittersweetdiabetes.com is the most popular blog.
- medtronicdiabetes.com, dexcom.com, novonordisk.com, lilly.com, myomnipod.com, tandemdiabetes.com are the main pharmaceutical companies in diabetes industry.
- jdrf.org is the most authoritative association.
- diabetesdaily.com is the mainstream media website talking about diabetes.
- childrenwithdiabetes.com is the largest diabetes online community.
- diabeteseducator.org is the biggest diabetes society.
- integrateddiabetes.com is one consulting institution which offer the integrated diabetes consulting services.

If we want to have some general ideas about which community mainly talking about which topic, we need to manually or semantically annotate each website using terminology of the diabetes fields would be a good way towards explanation the communities' contents.

In the next Chapter, we will explain how to use terminology approach to tag the 430 websites and explore further about the semantic meaning behind 5 discovered classes.

4 Chapter 4 Semantic Interpretation of the Map with Diabetes-Related Websites

In the previous chapter, we created and explored a network of 430 diabetes-related websites. With a combination of two state-of-art tools, Hyphe and Gephi, we respectively crawled and visualized this topic-sensitive network to reveal chronic diseases stakeholders' organizations and communities. The results show that diabetes online community has its own space. It is organized with visually detectable clusters and a peculiar distribution of websites. To further study this obviously non-random distribution, we applied a community detection algorithm aimed at detecting clusters of similarly connected nodes to provide a view at the sub-communities inside the diabetes community. In the end, we got a map-like network associated with 5 discovered clusters.

However, a question remains as to what these clusters refer to. Besides, if we are capable of finding a common reference, we need to push the question further and explain the relationships intra-network with a semantic analysis. This shall allow us to explain why some websites are closer than others from a network perspective. In short, we need to find the common reasonable explanation as to why websites are in a same cluster and why there is a difference between clusters while clusters are only a topological algorithmic construction.

4.1 Community detection

Community detection in complex networks has attracted a lot of attention in recent years (Newman, 2004). The main reason is that complex networks (Clauset et al., 2004) are made of a large number of nodes with hidden structures and most previous quantitative investigations focused on statistical properties disregarding the roles played by specific sub-networks. Detecting communities (or modules) can be a way to identify sub-structures which could correspond to important functions. This is the case with web networks for example, where communities are sets of web entities (pages and sites) dealing with the same topic (Boguna et al., 2003). Relevant community structures were also found in social networks (Fiedler, n.d.), (Pothen et al., 1990),

(Kernighan & Lin, 1970) the Internet (Zachary, 1977), food webs (Zhou, 2003), and in networks of sexual partners (Burt, 1976).

Detecting community structure is fundamental to uncover the links between structure and function in complex networks and yield many practical applications in many disciplines such as biology and sociology. A popular method now widely used relies on the optimization of a quantity called modularity (see Chapter 2), which is a quality index for a partition of a network into communities.

4.2 Semantic approach

As mentioned in Chapter 2, there are three main approaches to retrieve, extract and describe the semantic content of websites, terminology, ontology and Nature Language Processing. A modular ontology of the stakeholders doesn't yet exist in the context of the online diabetes field, and building such ontology is out of the scope of this work. Indeed, building such an ontology is an intensive work and is very time consuming in itself. Moreover, Natural Language Processing, at this early stage of our research work, was also considered resources and time consuming and we adopted first a terminology approach. We proposed a terminology to tag the websites.

Since our hypothesis is that a semantic description of the diabetes-related websites with tags could allow to predict the community clusters of the network, we need to define and organize the related tags to describe all content of 430 websites. In this study, we use thematic analysis which is one of the most common forms of analysis within qualitative research. It emphasizes identifying, analyzing and interpreting patterns of meaning (or "themes") within qualitative data. Thematic analysis is often understood as a method or technique in contrast to most other qualitative analytic approaches such as grounded theory, discourse analysis, narrative analysis or interpretative phenomenological analysis (Ibrahim, 2012). These approaches can be described as methodologies or theoretically informed frameworks for research as they specify guiding theory, appropriate research questions and methods of data collection, as well as procedures for conducting analysis. Thematic analysis is best thought of as an umbrella term for a variety of different approaches, rather than a singular method. In thematic analysis, little or no predetermined theory, structure or framework is used to analyze data; instead the actual data itself

is used to derive the structure of analysis. In this approach, the themes are strongly linked to the data since they emerge from it (Braun et al., 2014) (Guest et al., 2012).

If we want to semantically explain the clusters uncovered in the network, the content of 430 websites is data and we should analysis the data. Like most research methods, the process of data analysis can occur in two primary ways: inductively or deductively. In an inductive approach, the themes identified are strongly linked to the data. This means that the process of data analysis occurs without trying to fit the data into a pre-existing theory or framework (Thomas, 2003). However, it is important to note that induction in thematic analysis is not a "pure" induction free of all individual preconception as it is not possible for the researchers to free themselves from ontological, epistemological and paradigmatic assumptions. Hence, the analysis will always reflect the researcher's philosophical standpoint and research values.

Deductive approaches, on the other hand, are theory-driven. This form of analysis tends to be more interpretative because analysis is shaped and informed by pre-existing theory and concepts. Deductive approaches can involve seeking to identify themes identified in other research in the data-set or using existing theory as a lens through which to organize, code and interpret the data. A thematic analysis can also combine inductive and deductive approaches (Stacey, 2019).

The overarching purpose of this chapter is to investigate if there is a specific reason, in relation with semantic analysis, that explains why a website belongs to one class rather than another in the map obtained from our previous work in network analysis. In that case, we employ the inductive thematic analysis only to get the main idea of 430 diabetes-related websites' content. We set several categories of tags representing different dimensions of the websites' qualities and topics themselves representative of stakeholders. We tagged each website according to these categories and performed various analysis to better understand the relationship between tagging results and the clusters found by a community detection method. In the end, we apply auto correlation to present the most important tags for computers predicting the clusters.

4.3 Materials and methods

4.3.1 Materials

The material is a dataset including the 430 diabetes-related websites that were obtained in the previous study using a web crawler Hyphe. Each website is associated with the class it belongs to. Table 2 shows a sample of 10 websites with their belonging clusters from the raw dataset. Here we use class 1, class 2, class 3, class 4 and class 5 as notations to distinguish the classes.

Table 2. Sample of the dataset with 10 random websites and their belonging clusters.

Website	**Description**	**Class**
Diabeticinvestor.com	Diabetes Investor is the premier subscription-based content publisher that provides real time analysis of the business of diabetes.	1
Affordableinsulinproject.org	The Affordable Insulin Project offers tools, resources, and data so that people impacted by today's rising health care costs can positively influence the affordable access to this life essential drug.	1
Smashtastic.Wordpress.com	Smashtastic is the blog of one 27 years old woman with type1 diabetes.	2
Sweetlyvoiced.com	Sweetlyvoiced is a blog of one mother with type1 diabetes.	2
Stripsafely.com	StripSafely is a Diabetes Online Community (DOC) to help the general public understand that there are inaccurate blood glucose test strips and meters on the market.	3
Thetype2Experience.com	The Type 2 Experience is a collaboration blog created by a group of friends who live with type 2 at different levels and with different backgrounds.	3
Adorndesigns.com	The Adorndesigns is the online shop providing "High Style – Low Profile" handbags for diabetics.	4
Dexcom.com	Dexcom is a global Continuous Glucose Monitoring (CGM) System company.	4
Childrenwithtype1Diabetes.org	Childrenwithtype1Diabetes is an organization to provide support and information to parents of children diagnosed with Type 1 Diabetes.	5
Jdrf.org	JDRF is the leading global organization funding type 1 diabetes research.	5

4.3.2 Data Annotation

4.3.2.1 Inductive Thematic Analysis

We used inductive thematic analysis (ITA) to build categories to describe the content of the websites with different dimensions. A diabetes expert with 15 years type 1 diabetes experience proposed the initial set of categories to annotate the websites according to the stakeholders of diabetes, language and type of media in the diabetes-related websites, type of diabetes and diabetes-related topics, etc. Then, the project team reviewed a random sample of the websites and annotated each website inferring another set of categories. From these two propositions, we decided a final categorization for the tags selected for the annotation process.

4.3.2.2 Inter-rater Reliability

In order to check if the annotation process is annotator-dependent or not, we first scored the inter-rater reliability on a small number of websites. Inter-rater reliability measures the level of agreement among raters (Hallgren, 2012). We chose 2 individuals and reviewed 19 websites randomly selected from the 430 diabetes-related websites for inter-rater reliability of the tags. A simple percent agreement calculation (K. Krippendorff et al., 2011) was adopted to analyze the results of this preliminary process.

4.3.2.3 Annotation Process

The diabetes expert annotated manually all the 430 websites with the final tags. It took approximately one month to complete the whole annotation procedure.

4.3.2.4 Class Prediction

To study which combinations of tags can predict or explain the classes/communities/clusters according to our previous study, we used RapidMiner studio [6] framework to apply 7 different state-of-the-art clustering models to our dataset. The modeling methods are: Naive Bayes, Generalized Linear Model, Deep Learning, Decision Tree, Random

[6] Rapidminer Studio is a visual workflow designer software to import, prepare, and clean the data and then perform state of the art data science and machine learning algorithm.

Forest, Gradient Boosted Trees and Support Vector Machine (See table 3). For evaluation purposes, the dataset was split into a training set for the machine learning algorithms to learn and a testing set to determine the quality of the predictions. We used the 38 tags as features and the cluster ID as a target label to feed the different models. We chose a set of 60% random websites for training and 40% for testing and repeated several times the experiment. This ratio for training/testing could be higher but with the risk of overfitting the data. The ratio is also limited by the relatively small dataset used for the training.

Table 3. Seven different state of the art clustering models used by Rapidminer.

Method	Description
Naive Bayes	In machine learning, Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.
Generalized Linear Model	In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models' other than a normal distribution.
Deep Learning	Deep learning is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms. Deep Learning can be supervised, semi-supervised or unsupervised.
Decision Tree	A decision tree is a decision support approach that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.
Random Forest	Random forests or random decision forests are a family of learning methods for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.
Gradient Boosted Trees	Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
Support Vector Machine	In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data use for classification and regression analysis.

4.4 Results

4.4.1 Category Sample from ITA

At the beginning, the diabetes expert proposed the initial categories with 10 categories. The 10 categories are language, type of diabetes, URL, organizations, institutions, hospitals, people (who build the websites), people (targets), diabetes and the pharmaceutical industries. Each category has its own values (see table 3)

Table 4 The different values for 10 initial categories proposed by diabetes expert.

Categories	Values
Language	English, French, Chinese, Others
Type of diabetes	Type 1, Type 2, Gestational, Others
URL	Information, E-commercial, Blogs, Microblogs, Forums, Others
Organizations	International, Regional (National), Local
Institutions	Private, Public
Hospitals	Pediatric Hospitals, Public Hospitals, Non-profit Independent Hospitals, Private For=profit Hospitals
People (who build the websites)	Physicians, Endocrinologist, Diabetes specialists, Diabetes Educators, Dieticians, Nutritionist, Patients
People (targets)	Physicians, Patients (Children, Adults), Not specialized
Diabetes	Preventions, Diagnosis, Treatment (Drags, Devices), Follow-up, Well Being (Nutrition, Sports, Sleep, Tobacco, Alcohol)
Pharmaceutical Industries	Companies, Laboratories, Others

Then the project team improved it by using MindNode. MindNode is a visual brainstorming and mind mapping software that helps people connect their thoughts and clarify their ideas (Fursin et al., 2015). The second version of our categories is presented below (see figure 20). The different color presents different category and red represents the new propositions.

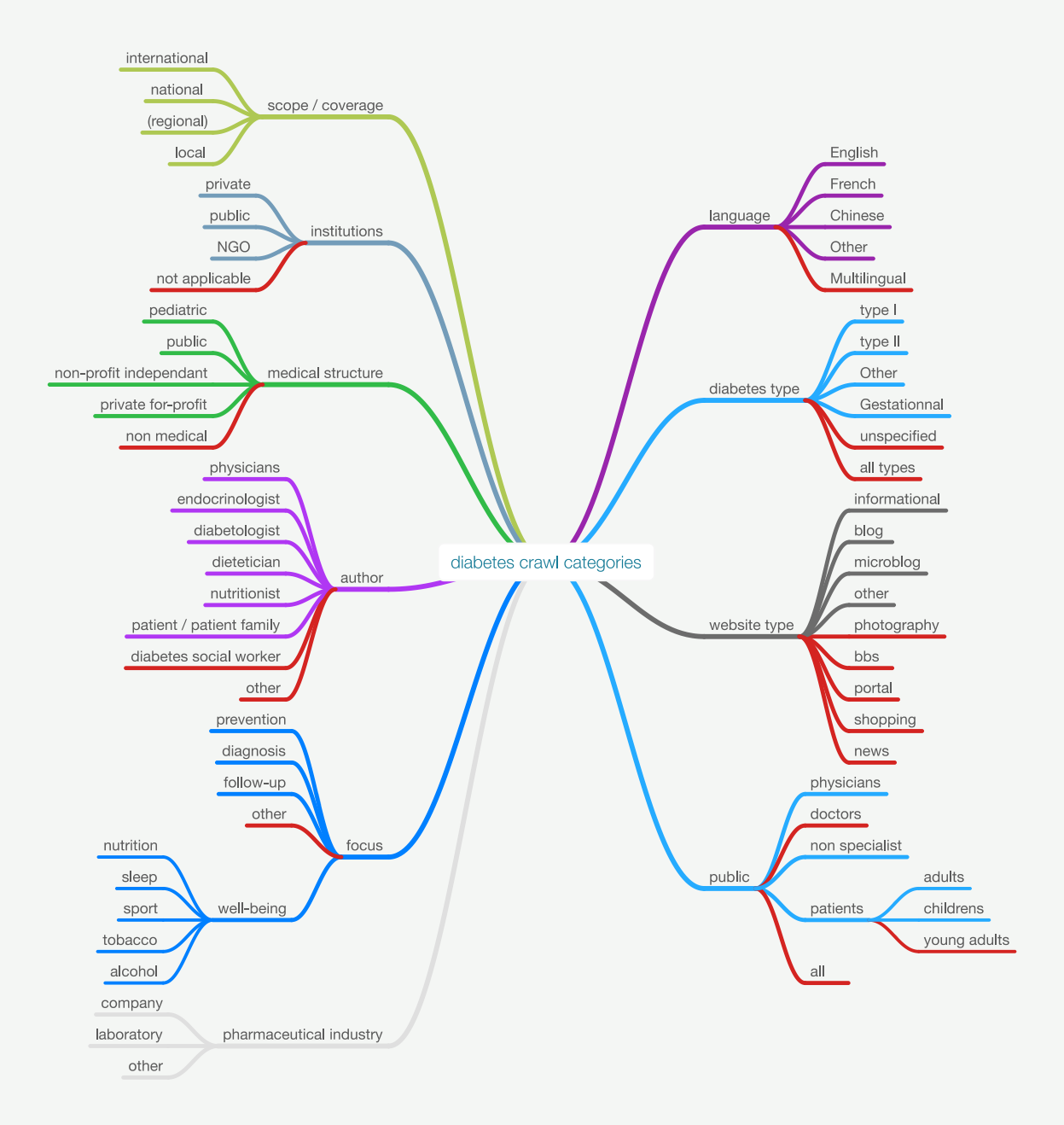


Figure 20 Diabetes crawl categories proposed by project team using MindNode.

In the end, we combined two propositions and considered the tags should be more simple but clear enough to describe the context of websites. Six categories with different values are identified: Status of stakeholders, Language of websites, Type of websites, Organizations, Type

of diabetes, and Diabetes-related topics. The 6 categories providing 38 values are presented in table 5. For some of them, the values are mutually exclusive as status, language, type of websites, organizations and for the others multiple values are authorized as type of diabetes and diabetes-related topics.

Table 5 The output of 6 categories with 38 values for tagging 430 diabetes-related websites.

Status of Stakeholders	Language of Websites	Type of Websites	Organizations	Type of Diabetes	Diabetes-Related Topics	Total
Non-Profit	English	Portal	Individual	Type 1	Prevention	
Profit	Multilingual	Information	Association	Type 2	Treatment	
		Blog	Society	Gestational	Self-management	
		Forum	Federation	Pre-diabetes	Advocacy	
		E-commerce	Charity		Complications	
		Click-to-donate	Company		Psychological Support	
			Program		Accessories	
			Conference		Sport	
			Hospital		Diabulimia[7]	
			Clinic			
			Pharmacy			
			Laboratory			
			Consulting			
			Media			
			Online Community			
2	2	6	15	4	9	38
Maximum number of possible tags from each category to annotate one web-site						
1	1	1	1	4	9	17

[7] Diabulimia is one food disorder related to diabetes.

In order to standardize the definitions of each tag, we use HeTOP as a tool to give some categories standard definitions. HeTOP contains the main terminologies or ontologies of the Health domain. It hosts more than 2 million concepts that are available in several languages among more than 70 terminologies or ontologies in 32 languages. It retrieves quality resourses by access powerful search engines (PubMed, liSSa, LILACS, Doc'CISMeF, etc.) (Julien et al., 2013).

Five main categories:

Status: Non-Profit, Profit

Non-Profit: refers to an entity to pursue a common not-for-profit goal, that is, to pursue a stated goal without the intention of distributing excess revenue to members or leaders.
Profit: refers to an entity to run some business for the profit purpose, usually refers to money.

Language: English, Multilingual

English: English Only.
Multilingual: refers to one web entity has more than one language.

Type of Website: Portal, Information, Blog, Forum, E-commerce, Click-to-donate

Portal: refers to a site that provides a starting point or a gateway to other resources on the Internet.
Information: refers to a site that provides information of institutions such as government, educational or nonprofit organizations.
Blog: refers to a site used to post online diaries which may include discussion forums to share the personal experiences and thoughts to others, including professional bloggers.
Forum: refers to a site where people can hold conversations in the form of posted messages.
E-commerce: refers to a site offering goods and services for online sale and enabling online transactions for such sales.
Click-to-donate: refers to a site that is designed to donate to charity simply by clicking on a button.

Organization: Individual, Association, Society, Federation, Charity, Company, Program, Hospital, Clinic, Pharmacy, Laboratory, Consulting, Media, Online Community, Conference

Individual: refers to one web entity is edited by one or several multi-authors, such as diabetes bloggers, etc.

Association: refers to a group of people organized for a joint purpose.

Society: refers to whose membership is limited to health-care professionals.

Federation: refers to one organization forming associations into a single group with centralized control.

Charity: refers to one social welfare organization with programs designed to assist individuals in need. Here is also including the foundation.

Company: refers to legal entities made up of an association of people, be they natural, legal, or a mixture of both, for carrying on a commercial or industrial enterprise. Here is implying business entities with an aim of gaining a profit.

Programme: refers to one research focusing on diabetes, usually is partnership with one cooperation or initiated by pharmaceutical company.

Conference: refers to the organization only for conference purpose.

Hospital: refers to institutions with an organized medical staff which provide medical care to patients.

Clinic: refers to the place where outpatients are provided medical treatment, checkup or advice for their health. It is frequently run by one or several general practitioners.

Pharmacy: refers to the place preparing and providing medications.

Laboratory: refers to facilities equipped to carry out investigative procedures.

Consulting: refers to one web entity is to offer the consulting service for diabetes.

Media: refers to one web entity is mainly to offer all publications on the website, including magazine, news, paper, journal, articles, etc.

Online community: refers to one web entity is a virtual community whose members interact with each other primarily via the Internet.

Types: Type1, Type2, Gestational, Prediabetes

Type 1: is an insulin-dependent diabetes.
Type 2: is a metabolic disease that affects insulin level within the body. The pancreas still produces insulin; however, the amount is insufficient.
Gestational: is a condition in which a woman without diabetes develops high blood sugar levels during pregnancy.
Prediabetes: is the precursor stage before diabetes mellitus in which not all of the symptoms required to diagnose diabetes are present, but blood sugar is abnormally high.

Diabetes-related topics: Prevention, Treatment, Self-management, Advocacy, Complications, Psychological Support, Sport, Diabulimia, Accessories, Health-care Professionals

Prevention: mainly talking about before diagnosis diabetes.
Treatment: mainly talking about the medications, devices, suppliers, etc.
Self-management: mainly talking about the care after the diagnosis from patients themselves and also from health care professional side, including healthy diet and exercises. Self-management is somehow equal diabetes education.
Advocacy: mainly doing the activities to improve the social awareness about diabetes.
Complications: mainly talking about the eyes damage, kidney failure, feet problems, etc.
Psychological Support: mainly talking about one's psychological development in, and interaction with, a social environment.
Sport: mainly talking about the fitness and life activities.
Diabulimia: one food disorder related to diabetes.
Accessories: selling the accessories of diabetes suppliers, such as diabetic insulin case, pump stickers, or diabetic necklace, T-shirts, mugs, etc.
Health-care Professionals: mainly refer to the websites of health-care professionals' society

4.4.2 Inter-Rater Reliability

To test tags validity, categories and description, we proceeded to test the annotation process. The goal was to determine if the proposed tags are univocal, described enough and sufficient to allow (a) a straightforward decision-making when tagging and (b) a replicability of the study with the same data. To achieve this goal, we chose to proceed with an inter-rater reliability test consisting into asking a set of raters to tags a sample of websites and measure their agreement. As this task is a simple validation process for a set of tags that will ultimately be used by one rater only, we chose not to go full length with the inter-rater scoring involving a more reliable inter-rating reliability test such as Cohen's Kappa (J. CohenA et al.,1960) and larger sample but instead chose a reasonably small amount of websites to proceed to a Joint-Probability of Agreement test with 3 raters (Uebersax, 1987). Again, the goal here was to determine how practical the methodology of tagging really is and if other teams would be able to replicate the study a minimal.

We chose a small random sample of 19 websites accounting to 4.4% of the corpus to keep a tagging time under 24h. Initially we chose 20 websites but one website was not accessible when the ranking took place so it was removed. Each tag has to be chosen by reading the website and deciding on the tag. This process can take up to 5min per tag. The websites were tagged at the same time and date from the same physical location (see chapter 2 for web precautions). Since 4 categories such as Status of Stakeholders, Language of Websites, Type of Websites and Organizations are mutually exclusive and the rest of categories such as Type of Diabetes (4) and Diabetes-Related Topics (9) have binominal values, we presented 17 (4+4+9=17) tags in our inter-rater reliability result (see the last 2 rows in table 3). After the annotation process by the 3 raters, each rater performs 323 evaluations (17*19=323). A value count of the raters' agreement (0, 2 or 3) was assigned to each evaluation according to the number of raters that agree for the evaluation.

The figure 21 presents the final results of agreements on 17 values with the 19 websites. Among them, there are 212 evaluations with a score of 3, 111 evaluations with a score of 2 and no evaluation with a null score. If 3 raters would agree on all evaluations, the highest score possible is 969. There is a 65.63% of unanimous agreements. Moreover, with 212 agreements with score 3

and 111 agreements with score 2, the total score is 858 which hits 88.54% accuracy. This inter-rater reliability score is not perfect but lead to a discussion with the raters to share the tagging experience. It was decided that the agreement is high enough to continue the study with a minor change to the tags and a better description. This allowed an acceptable replicability and a forthright decision-making.

The change concerned the 4 tags “profit”, “non-profit” (Status of Stakeholders) “English” and “Multilingual” (Language of Websites) that were reduce to 2 tags: “Profit” (if the tag is absent it means that it is a non-profit website) and “English” (if this tag is absent it means that it is a multilingual website).

Figure 21 Agreements on 17 values with the 19 diabetes-related websites.

4.4.3 Annotation Process and Tags Distribution

Using the set of tags now established, the whole corpus (430 websites) have been tagged with the 36 different tags.

To illustrate the result of this process, table 5 shows the tags that have assigned to the 10 websites presented in table 6.

Table 6 The output of the 10 sample websites with their tags and each cluster.

Website	Tags	Class
Diabeticinvestor.com	Profit, English, Information, Consulting, Type1, Type2, Gestational	1
Affordableinsulinproject.org	English, Information, Association, Type1, Type2, Gestational, Treatment, Advocacy	1
Smashtastic.Wordpress.com	English, Blog, Individual, Type1, Self-management, Psychological Support	2
Sweetlyvoiced.com	English, Blog, Individual, Type1, Self-management, Psychological Support	2
Stripsafely.com	English, Information, Online Community, Type1, Type2, Gestational, Advocacy	3
Thetype2Experience.com	English, Information, Media, Type2, Self-management, Complications	3
Adorndesigns.com	Profit, English, E-commerce, Company, Type1, Type2, Gestational, Accessories	4
Dexcom.com	Profit, Information, Company, Type1, Type2, Gestational, Treatment	4
Childrenwithtype1Diabetes.org	English, Information, Online Community, Type1, Self-management	5
Jdrf.org	English, Information, Association, Type1, Treatment, Self-management, Advocacy, Psychological Support	5

The 36 tags are unequally spread among the websites. As shown on figure 22, some tags are almost ever present while others are very specific and almost never used. However, each tag has been at least tagged once and no tag covers the whole corpus. On average, a tag is used 80.8 times but the count follows a power law[8] distribution and tags like diabetes type 1, language (English), self-management and website type information are widely found in more than half of

[8] In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independent of the initial size of those quantities: one quantity varies as a power of another.

the website (count > 215). They describe the corpus as a whole and define the broader topic this corpus is about: diabetes type 1, in English providing information about how to manage one's diabetes. In the other hand, 20 tags are used in less than 10% of the websites and are either very specific to a niche-like topic as diabulimia (Diabulimiahelpline.org) or simply non-relevant at all, for example, some organization type like consulting (Diabeticinvestor.com) which is giving the information to people who need to develop their business related to diabetes or pharmacy (Diabetesexpress.ca).

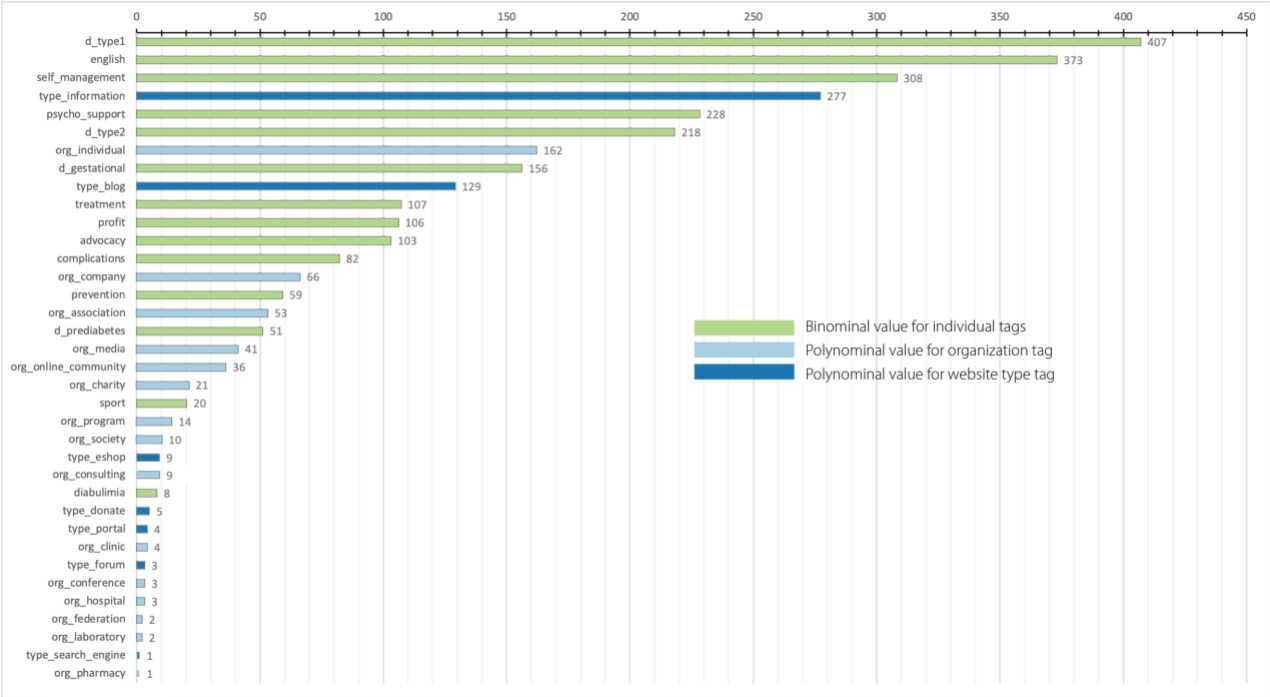


Figure 22 Tags count on the 430 websites (The tag names are prefixed with their abbreviated category).

4.4.4 Tags Analysis

The tags goal was dual: on one hand, we needed to qualify the whole corpus topics to provide general information about it, and some tags were useful for that; on the other hand, we

also used the tags as a discriminating mechanism to identify regions or localities inside the corpus if they existed.

To better understand how tags can help analyze the corpus and which tags are best for this job, we used two methods that yields similar result: principal component analysis (PCA) and weight by tree importance. In both methods, the goal is to quantify the quality of the tags to explain the corpus.

The objective of the PCA is to reduce the set of 36 individual tags to a much smaller set while still preserving the variance so as to not loose information.

To preserve 95% of variance, the PCA needs to keep 18 components as seen on figure 23. This shows that no simple component can explain a lot a variance and that the corresponding tags combinations hardly explain the 430 websites tags signature. That being said, the tags that contribute the most in the components' variance, are always the same ones. We created an index of the tag contribution by (1) taking the first 15 components ranked by their contribution to variance and by (2) calculating the average contribution of each tag to the component multiplied by the component contribution to the global variance. The result is the importance of each tag, according to the PCA to explain the corpus. The result is show in figure 24.

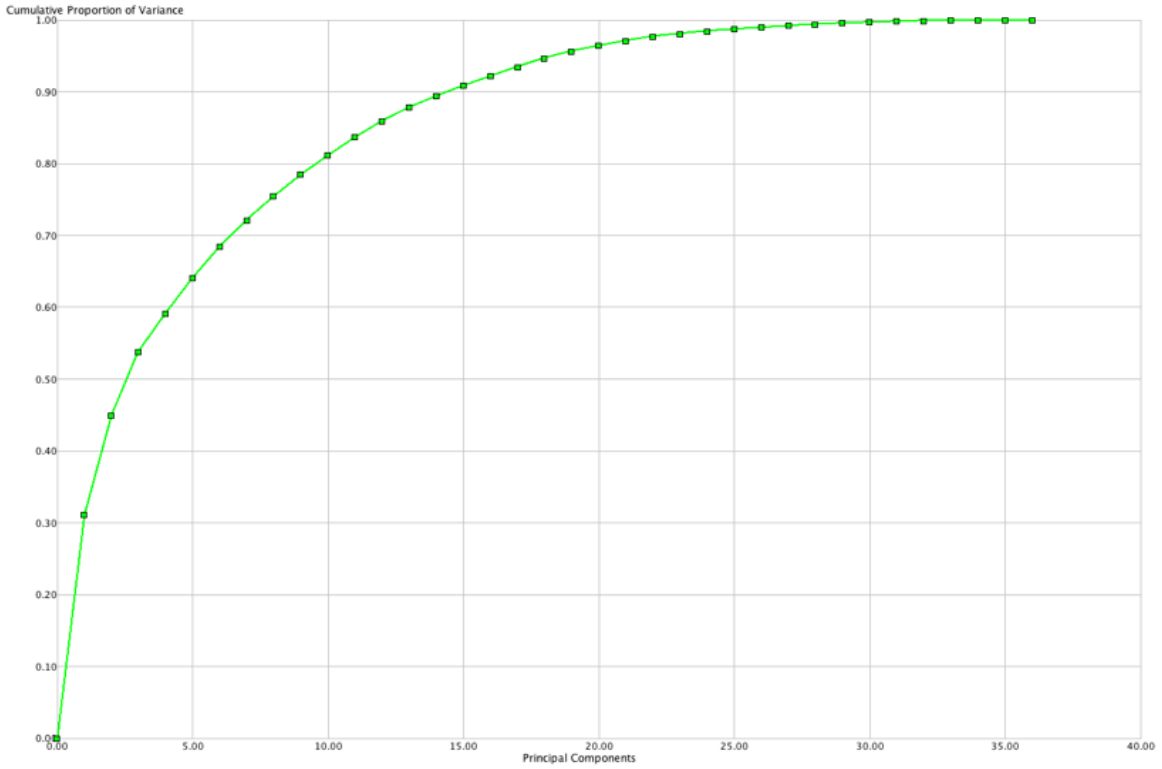


Figure 23 Cumulative Variance of Principal Components on 36 Tags Values.

attribute		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
d_type2	0,016	0,398	-0,155	0,243	-0,234	-0,132	0,2	0,393	-0,31	0,276	-0,277	0,018	-0,021	0,297	0,036	0,017
type_information	0,015	0,367	-0,126	-0,427	0,02	-0,362	-0,211	-0,135	-0,064	-0,139	0,002	-0,048	0,042	0,009	0,115	-0,338
d_gestational	0,015	0,366	-0,123	0,356	-0,216	0,005	0,352	-0,037	-0,126	-0,13	0,408	0,163	0,226	-0,293	-0,02	-0,14
treatment	0,013	0,275	0,074	0,08	-0,21	0,496	-0,437	-0,412	-0,216	-0,02	-0,013	0,333	-0,144	0,132	0,143	0,034
profit	0,013	0,268	0,327	0,175	-0,121	-0,249	-0,185	0,131	0,511	-0,054	-0,053	0,141	-0,026	-0,067	0,053	-0,005
org_company	0,010	0,184	0,351	0,106	-0,12	-0,06	0,03	-0,006	0,353	-0,025	-0,101	0,048	-0,06	0,058	-0,018	0,204
complications	0,012	0,129	-0,407	0,172	0,091	0,157	-0,181	0,174	0,085	-0,111	-0,359	-0,124	-0,46	-0,509	-0,029	-0,002
d_prediabetes	0,009	0,113	-0,299	0,183	0,145	0,133	-0,01	-0,04	0,305	-0,129	0,26	-0,153	-0,052	0,257	0,107	0,078
prevention	0,010	0,112	-0,352	0,145	0,141	0,111	0,028	-0,144	0,332	-0,041	0,109	-0,265	0,046	0,402	-0,051	-0,081
org_media	0,008	0,071	-0,153	0,085	0,098	-0,026	-0,407	0,109	-0,174	-0,193	-0,109	-0,002	0,528	0,065	-0,425	0,3
org_association	0,009	0,042	-0,177	-0,143	0,1	-0,119	0,506	-0,427	-0,003	-0,182	-0,308	0,309	-0,162	0,086	-0,187	0,214
org_society	0,002	0,016	-0,032	0,011	0,051	-0,002	0,014	0,039	-0,043	0,01	0,007	-0,076	-0,076	-0,066	0,094	-0,199
org_charity	0,004	0,011	0,003	-0,08	0,063	0,121	0,077	0,104	-0,138	0,026	0,065	-0,228	0,121	-0,062	0,621	0,413
type_eshop	0,002	0,009	0,036	0,029	0,03	0,018	0,059	0,072	0,132	-0,016	0,014	0,044	-0,038	0,013	-0,07	0,268
org_program	0,002	0,009	-0,017	-0,069	0,031	0,034	-0,003	0,036	0,023	0,015	0,013	-0,059	0,071	0,045	0,073	-0,427
org_hospital	0,001	0,009	0,003	0,004	0,004	0,021	-0,015	0,005	-0,026	-0,011	-0,029	-0,005	-0,055	-0,06	0,003	-0,039
org_online_community	0,006	0,009	-0,053	-0,148	-0,163	-0,113	-0,119	0,088	-0,11	0,305	0,522	-0,056	-0,48	0,077	-0,372	0,183
org_conference	0,001	0,008	0,008	0,006	-0,012	-0,008	0	0	-0,015	-0,007	0,013	0,011	0,01	-0,011	0,021	-0,041
org_clinic	0,001	0,007	0,002	-0,004	-0,003	0,01	-0,019	-0,003	-0,018	0,001	-0,017	0,012	-0,026	-0,047	0,03	-0,039
type_portal	0,001	0,006	0,008	0,014	-0,003	0,019	0,053	-0,003	-0,044	-0,003	0,005	0,02	0,002	-0,001	-0,003	0,089
org_federation	0,001	0,004	-0,016	0,005	-0,023	0,019	-0,007	-0,014	0,017	-0,006	0,004	-0,016	-0,01	0,014	0,034	-0,014
org_pharmacy	0,000	0,004	0,004	0,002	-0,004	0,003	0	-0,008	-0,007	0,004	0,001	-0,007	0,004	-0,007	-0,002	-0,012
org_laboratory	0,000	0,004	0,001	-0,003	0,013	0,021	-0,009	-0,006	-0,02	-0,018	0,009	0	-0,025	-0,004	0,008	-0,019
sport	0,003	0,002	-0,041	-0,069	-0,06	0,034	-0,012	0,101	0,049	-0,103	-0,013	-0,105	0,133	-0,229	-0,263	0,02
type_forum	0,001	0,001	-0,011	0,011	-0,017	0,014	-0,006	0,015	-0,019	0,033	0,044	0,012	-0,077	0,006	-0,029	0,07
type_search_engine	0,000	0,001	-0,008	0,005	-0,011	0,016	0	0,003	0,011	-0,008	0,004	0	-0,004	0,001	0,001	0,012
org_consulting	0,000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
type_blog	0,013	-0,378	0,092	0,387	-0,036	0,24	0,063	-0,005	0,009	0,139	-0,081	0,065	0,052	0,035	-0,127	-0,316
psycho_support	0,015	-0,358	-0,145	0,109	-0,666	-0,245	-0,085	-0,03	-0,058	-0,484	-0,047	-0,156	-0,113	0,153	0,128	0,031
self_management	0,015	-0,226	-0,399	0,178	-0,016	-0,398	-0,225	-0,208	0,149	0,478	-0,004	0,336	0,152	-0,125	0,226	0,096
language	0,009	-0,142	-0,107	-0,051	0,265	0,006	-0,046	0,493	-0,02	-0,36	0,191	0,61	-0,129	0,172	0,141	-0,047
d_type1	0,003	-0,017	-0,03	0,016	-0,035	0,003	0,055	-0,138	0,036	-0,152	0,313	0,038	0,092	-0,388	0,01	0,094
advocacy	0,011	-0,007	-0,257	-0,486	-0,448	0,399	0,116	0,212	0,353	0,183	-0,064	0,174	0,224	-0,042	-0,035	-0,012
diabulimia	0,001	-0,003	-0,024	-0,006	-0,019	0,03	0,007	-0,032	0,021	-0,083	-0,006	0,014	-0,02	0,044	-0,001	0,028
type_donate	0,001	-0,002	0,01	-0,015	0,023	0,054	0,042	0,051	-0,03	-0,004	0,007	-0,095	0,019	-0,059	0,103	0,189
org_individual		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 24. Tags ranking according to their PCA component contribution.

The top 10 tags contributing the most to the corpus variance are: d_type2, type_information, d_gestational, treatment, profit, complications, type_blog, psycho_support, self_management, advocacy. These tags are not the one with higher count but carry the more interest to help identify the 430 websites.

We applied a different method to assess the tags importance with the weight by tree importance, a derivative method from random forest[9] (Menze et al., 2009). This method calculates an information gain by splitting the corpus according to each tag to create 100 decision trees of depth 10. These trees are then compared and yielded a tag weight according to their contribution to the potential classifications. The particularity of this method is to also consider outliers websites whose tags behavior is peculiar where the PCA would ignore them.

As shown in figure 25, the tags are quite similar with the PCA results with some differences reflecting the outliers importance for the trees to take everything into consideration and not the most representative ones. Therefore, some tags whose count is very large (d_type1) or very small (diabulimia) can help the trees to take a decision while PCA would undermine their usefulness.

[9] Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

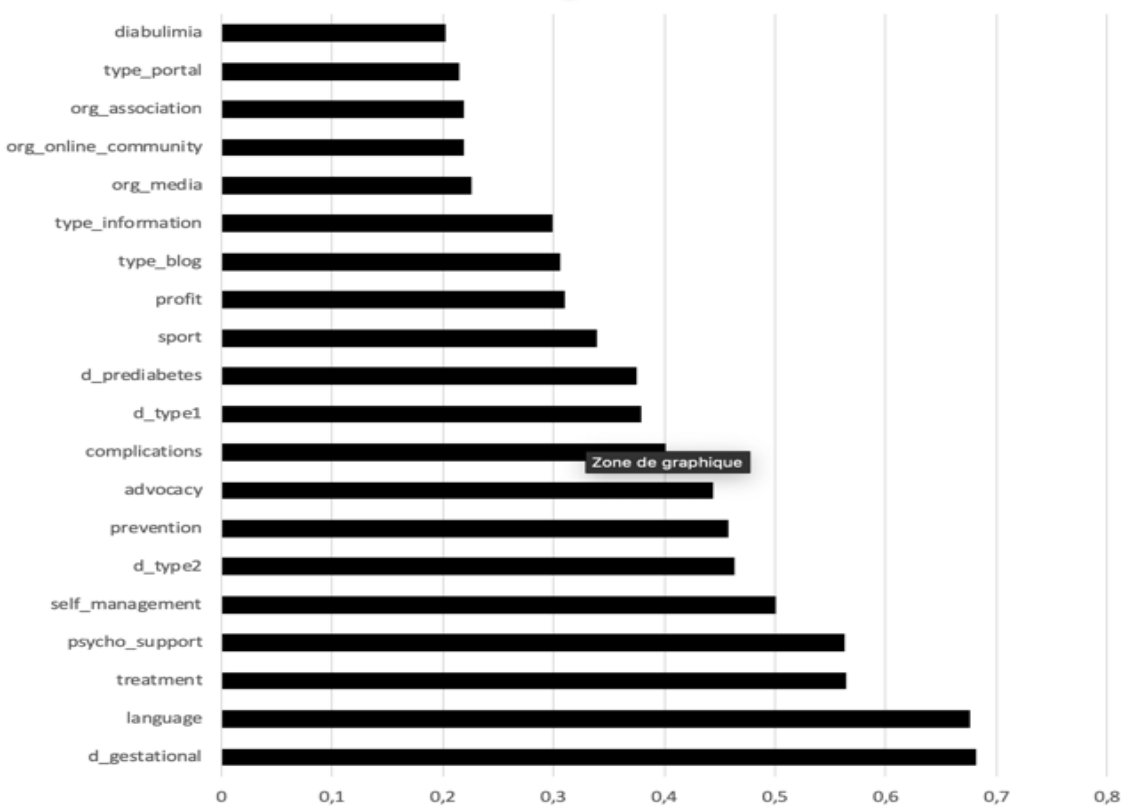


Figure 25. Tags weight ranked by importance for random forest modeling.

To conclude on the tags importance, both methods yield a slightly different set of tags. PCA and Random Forest both highlight different characteristics in tags. Hence, they helped assess which tags are indeed useful to discriminate the websites and which tags are not.

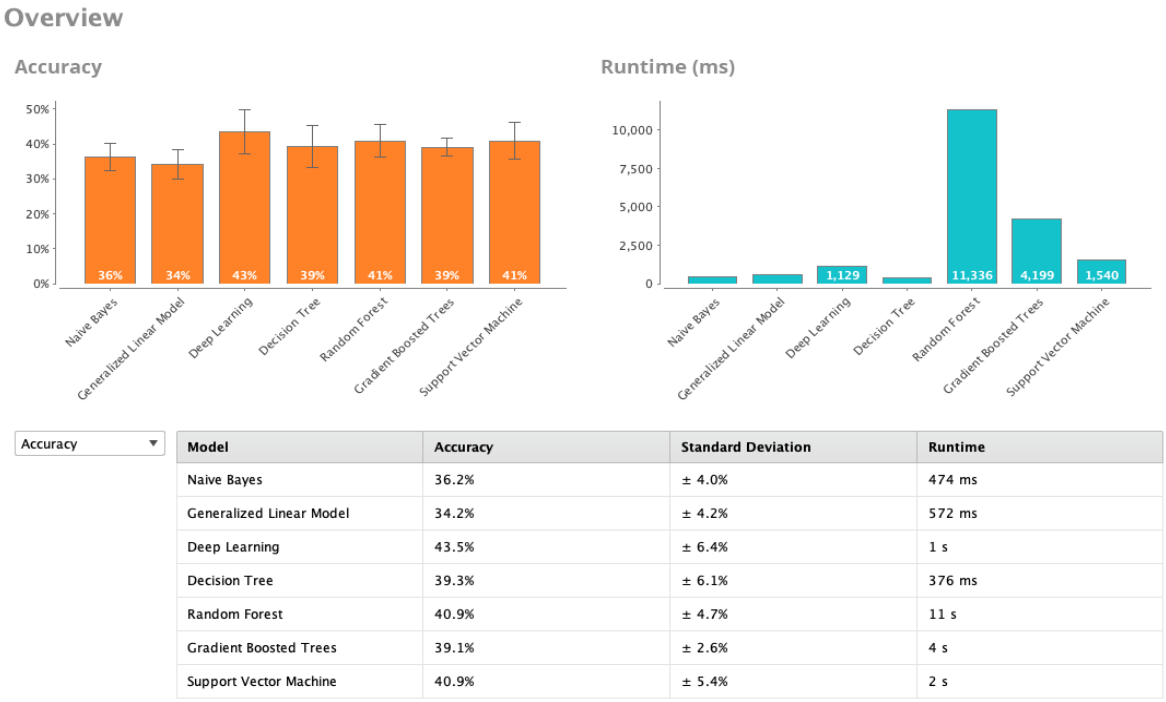
4.4.5 Classes Prediction

Our second goal was to determine if the content of website type alone described with tags represented by the tag distribution was enough to determine the clusters of links of these websites as found in prior work.

The seven machine learning models were applied to the data obtained by the annotation process to predict clusters according to tags. The global performances and runtime to predict clusters from tags are present in figure 26. The performance refers how many times machine can predict correctly the website's class from the tags model. Precision is defined as the number of true positives over the number of positives plus the number of false positives. Recall is defined as

the number of true positives over the number of true positives plus the number of false negatives. Recall and Precision are sometimes combined into a f-score but we chose to keep both information here as results are not exactly good.

A system with high recall but low precision returns many results, but most of its predicted labels are incorrect when compared to the training labels. A system with high precision but low recall is just the opposite, returning very few results, but most of its predicted labels are correct when compared to the training labels. An ideal system with high precision and high recall will



Model	Accuracy	Standard Deviation	Runtime
Naive Bayes	36.2%	± 4.0%	474 ms
Generalized Linear Model	34.2%	± 4.2%	572 ms
Deep Learning	43.5%	± 6.4%	1 s
Decision Tree	39.3%	± 6.1%	376 ms
Random Forest	40.9%	± 4.7%	11 s
Gradient Boosted Trees	39.1%	± 2.6%	4 s
Support Vector Machine	40.9%	± 5.4%	2 s

Figure 26 The global performances of 7 models and runtime to predict clusters from tags.

return many results, with all results labelled correctly. However, unfortunately, no model stands out as a good predictor of our dataset.

The accuracy of the 7 models is 40% at best. This indicates a weak prediction capability. Indeed, no model really stands out indicating that the cause of the weakness shall probably lie in the data as all the models use the same data as an entry point. To get into more details about the performance of the model, we chose the relatively high accuracy model, random forest which generates a forest of decision trees of variable size and depth. The optimal parameter for the

random forest is 140 trees with a maximal depth of two, which is an average parameter setting for such a dataset property (in terms of data count and data structure).

The detail accuracy of the random forest model on figure 27 shows two distinct phenomena. On one hand, the model can predict something when websites belong to cluster 2 or cluster 1 (the two most populated clusters) even if its accuracy in doing so is very low. Actually, the class recall of class 1 is 80% while the class recall of class 2 is almost 70%, indicating that the model returns actual predictions for these classes; however, their low-class precision indicates that even though we can predict something for them, making a true prediction is still hard to do with 42% precision average, (35.94+48.33)/2 (see figure 27). To summarize, many results are returned but most of the prediction are incorrect. One the other hand, this is not true at all for the 3 other classes. As a matter of fact, none of the 3 other clusters (class 3,4,5) have been predicted by the random forest model. This indicates that the tags we created worked on two clusters only but are not specific enough to really predict these two clusters while the 3 remaining clusters are simply not represented by the tags. The latter are indeed too heterogeneous in their tags' distribution and not specific enough for a group of tags to be predicted accurately.

accuracy: 41.93% +/- 4.00% (micro average: 41.94%)

	true class2	true class4	true class1	true class3	true class5	class precision
pred. class2	23	7	7	20	7	35.94%
pred. class4	0	0	0	0	0	0.00%
pred. class1	10	9	29	3	9	48.33%
pred. class3	0	0	0	0	0	0.00%
pred. class5	0	0	0	0	0	0.00%
class recall	69.70%	0.00%	80.56%	0.00%	0.00%	

Figure 27 Random forest model performance for predicting each cluster according to the tags.

In addition, to better understand how tags can help analyze the corpus and which tags are best for predicting clusters, it is possible to determine the weight of each tag as the global importance of each of the original tag or category for each of the corpus cluster, independently of the modeling algorithms. When calculating such a weight, as shown in figure 28, we got the

following top 10 weights (weight has been rescaled and normalized between 0 and 1): WebsiteType blog (1.0), WebsiteType information (0.90), Psychosocial Support = true (0.63), Gestational = false (0.55), Type2Diabetes = true (0.55), Treatment = false (0.41), Complications = false (0.41), Complications = true (0.40), Language = English (0.34), Prevention = false (0.32). High weight tags are the most useful tags used in the 7 models and their distribution is localized more in better performing clusters 1 and 2 than the 3 others.

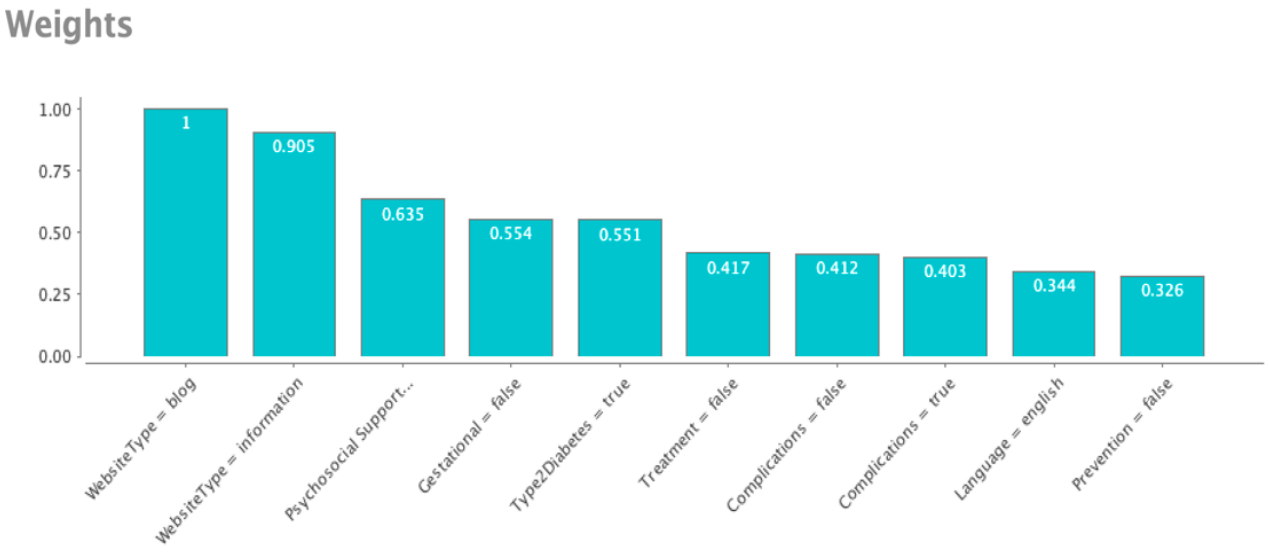


Figure 28 The top 10 tags contributing the most for predicting clusters.

As a conclusion for the measure of the tags' importance, some tags are indeed useful to explain the corpus variety of websites tags distribution for two clusters representing half of the corpus while the other tags are either too general and qualify the corpus as a whole but not individual clusters or too specific and not representative of a cluster (size > 50 websites) but rather of a very small group of websites (between 1 and 10 websites in the group).

Results show a low prediction performance by using tags to provide a semantic explanation of the clusters of DiaMap, at least two clusters are clearly defined by a few specific tags and the others are mixed. Without the successful predicting performances, we better understand the tags distributions and their relative importance. This can help us to refine the tags for future broader analysis of the diabetes web space. The result of the clusters prediction also reflects the community

reality: a mix of websites of different types that create a mixed but localized space. It proves the community has a tagging scheme sometimes, but it is still hard to use semantical approach to predict accurately the clusters. In practice, this work will allow us to improve the search engines to find more relevant and accurate information by using semantic tags.

In the next chapter, we will explain how DiaMap works in the context of information retrieval.

5 Chapter 5 Accessing DiaMap for Accurate Information Retrieval and Domain Awareness

In this Chapter, we will use several practical scenarios to explain how DiaMap works in the context of information retrieval. We will design a protocol to compare information retrieval quality between DiaMap and usual search engines.

Five questions are proposed and asked to four chosen search engines to find the websites which can answer the questions properly. Then we will perform 20 experiments with this protocol. After all, we will set a metric to compare the results offered by the search engines and DiaMap and figure out how users can get what they want in certain time and how satisfied they are with the results presented by the list of websites and the graphs.

5.1 Introduction of Selected 4 Search Engines

Online information targeting chronic diseases allow users gain simplified access to the Web to find online resources and obtain health information or social support (Purcell et al., n.d.). The most prevalent entry points for such online health information are general search engines, which propose supposedly relevant websites through results pages (SERPs) from a sentence or set of words known as a *query*. (Purcell et al., n.d.) For a given query, hundreds of thousands of results are provided, and paginated into SERPs. This provides users with a major opportunity to find relevant information, but the full potential of the system is rarely used, as most users focus on the first three to five results on the first page and ignore the other pages. (Kim et al., 2015), (Cutrell & Guan, 2007), (Höchstötter & Lewandowski, 2009), (Granka et al., n.d.).

The main purpose is to assess the added value of querying DiaMap over the use of regular general search engines, we decided to use the most popular general search engines, to mimic non-expert users. These search engines are the most likely to be used for comprehensive online health information retrieval as they are both well-known and regularly used by all online users (NW et al., 2012). We selected the search engines for this study from the 10 most population search engines in 2019.

Many reports comparing the popularity of current search engines are available on the Internet, all reporting similar and consistent statistics. According to the website oberlo.com (*https://www.oberlo.com/blog/top-search-engines-world*) which identified the top 10 general search engines worldwide in 2019, the most popular search engine is Google (google.com), which currently holds the largest share of the search engine market worldwide. Google is so popular that its market share is many times larger than those of all the world's other search engines combined. Moreover, according to statistics from Netmarketshare [10] (netmarketshare.com), Statista [11] (statista.com) and Statcounter[12] (gs.statecounter.com), the top six search engines worldwide, in terms of market share, are: Google (81–93%); Bing (2–5%); Baidu (1–9%); Yahoo (1–2%); Yandex (0–1%) and Duckduckgo (0–1%) [*https://www.reliablesoft.net/top-10-search-engines-in-the-world. Accessed by February 17, 2020*]. For this study, we selected Google, Bing, Baidu and Yahoo as the general search engines of choice for the five experiments.

Baidu is relatively unknown in western countries, but is popular in China[13], serving 850 million netizens (China Internet Network Information Center., 2019). Baidu's results for queries in English are of lower quality than for queries in Mandarin, but we nevertheless included this search engine in this study because of the large number of users it represents worldwide.

For the purpose of these experiments, the query corresponding to the question had to be input with depersonalization, to prevent SERPs with customized results based on previous searches or user profiling being obtained (Shen et al., 2007). Queries were therefore performed in Paris, France, on the global, non-localized, search engine website.[14] The web browser was used in *in incognito* (or privacy) mode, with no cache and no login profile.

[10] NetMarketShare provides web usage share statistics on real users, trusted by Roche, CNN, Microsoft, Apple, SONY, Forbes, etc.
[11] Statista is a leading provider of market and consumer data. Over 600 visionaries, experts and users continuously reinvent Statista, resulting in the continual development of successful new products and business models.
[12] Statcounter simplifies website analysis and is trusted on over 2 million websites.
[13] https://www.iresearch.com.cn/coredata/2011q3_5.shtml
[14] Most search engines have a local version to serve netizens in their native language.

5.2 Scenario Applications

We designed a protocol for comparing information retrieval quality between DiaMap and a search engine. Then we performed for each application with this protocol, all of which were analyzed together.

5.2.1 The Protocol

The comparison protocol was designed to compare similar querying processes for DiaMap and a general search engine (SE), despite differences in the query interface. Figure 29 illustrates the protocol and the two processes generating comparable results.

The process starts (0) with a single question **Q** for a real-life question which user might ask about diabetes. The first step (1) is the translation of the question into a set of corresponding tags (T). The tags are selected from the 38 values presented at the end of section 1 (page 4). For the purposes of this study and the experiments performed, this step was performed manually, but it could be automated. The second step (2) is to query the SE with Q. This generates a list of websites **{W–SE}** from each search engine corresponding to the top five websites proposed. This reflects typical user behavior concerning SERPs. The third step (3) is to query DiaMap with T. This query generates a list of websites **{W–DiaMap}** corresponding to the tags and a **submap** of websites corresponding to a subgroup of DiaMap websites with at least one link in common with the websites in {W–DiaMap}. For DiaMap, we did not limit the number of websites selected as we did for the search engines, because the submap can display all the websites together, with fairly even weightings, in a single graphic.

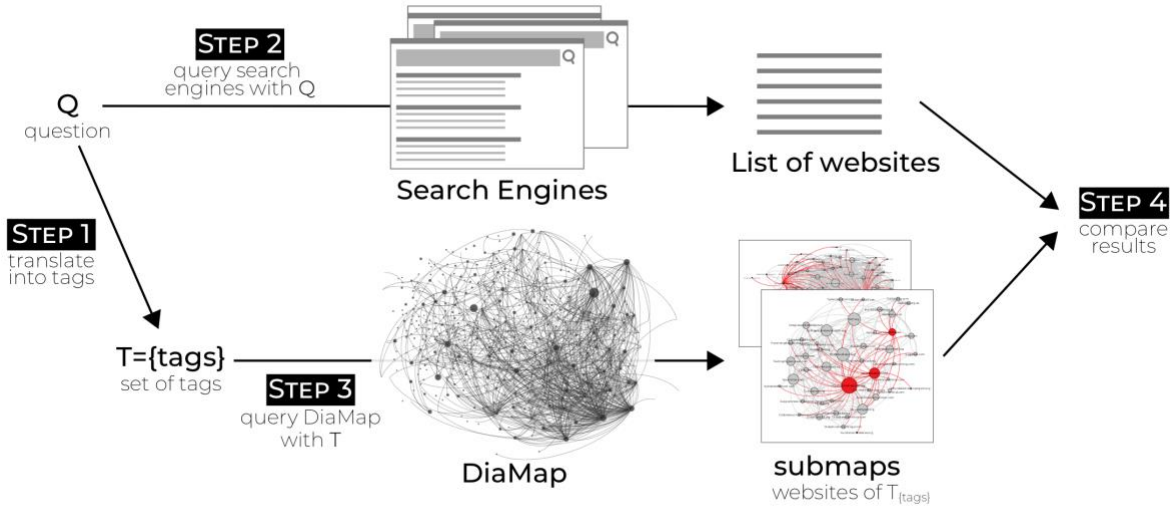


Figure 29 Illustration of the protocol for comparing DiaMap with a search engine, and the workflows to be compared.

In Step (4), the aim is to compare the results between DiaMap, consisting of {W–DiaMap} and the submap, with those generated by search engines {W–SE}. For this purpose, we analyzed {W-SE} further, by defining a relevant set of results (**RRSE)** and generating five numerical indicators: **RRSEcount**, **WDiaMapCount**, **RRSEintersectWDiaMap**, **SubMapCount**, **RRSEinSubMap**.

- **RRSE** (Relevant Results Search Engine): relevant results among the first five websites suggested by the SE in {W–SE}. This information was evaluated by a diabetes expert, who attributed a binary (*Yes/No*) value to each result. The diabetes expert assessed website relevance by going to the website concerned and reading its entire contents. The diabetes expert then assigned a *Yes* (relevant) or *No* (not relevant) value to each website. Only websites with a value of *Yes* were included in the RRSE.
- **RRSEcount** (Relevant Results Search Engine Count): cardinality of RRSE. RRSEcount was determined as the sum of {W–SE} with a relevance value of *Yes*. Its maximum value was 5 as {W-SE} itself yielded five results, so RRSEcount ∈[0–5].
- **WDiaMapCount** (Websites DiaMap Count): cardinality of the websites selected in DiaMap with the question Q translated in tags T {W-DiaMap}. It was not necessary to

determine the relevance of the websites in {W–DiaMap}, because the website tags included had already been determined by a similar evaluation method. All websites in DiaMap are, in fact, relevant. As DiaMap contains 430 websites, WDiaMapCount ∈ [0–430].

- **RRSEintersectWDiaMap** (Relevant Results Search Engine Intersect Website DiaMap): cardinality of {W-SE | relevant = yes} ∩ {W-DiaMap}. This indicator counts the number of websites present in both RRSE and {W-DiaMap}. This value cannot exceed the minimum of RRSEcount and WDiaMapCount so RRSEintersectWDiaMap ≤ min (RRSEcount, WDiaMapCount).
- **SubMapCount** ({W-DiaMap} hyperlinks submap count): cardinality of all the websites within DiaMap sharing with at least one hyperlink in common with each of the {W-DiaMap} websites. This indicator makes use of the densely connected nature of web networks. It provides an overview of the domain-specific surroundings of the {W-DiaMap} websites that the user might encounter. The websites in the submap are, thus, not immediately relevant to Q, but are closely related to it. From the user's perspective, the submap provides an understanding of the scope of the question and of the related domain-specific topics in the close context/neighborhood. Cardinality itself provides an indication of the extent to which the submap is populated, whereas the submap figure provides an indication of its complexity. As DiaMap contains 430 websites, the submap website count cannot exceed this figure, so SubMapCount ∈ [0–430].
- **RRSEinSubMap** (Relevant Results Search Engine in {W-DiaMap} hyperlinks submap): cardinality of the RRSE websites not directly included in {W-DiaMap}in the submap. This metric is complementary to RRSEintersectWDiaMap. It is less restrictive and expands the scope whilst remaining in the same immediate surroundings, in the diabetes domain. This indicator shows how DiaMap can offer users a greater variety of high-quality health information closely related to their topic of interest than general search engines. Its maximum value is 5, as {W-SE} yields five results, so RRSEinSubMap ∈ [0–5].

At the end of the process, each comparison yields the following quintuplet result for each question with each search engine Rij = (RRSEcount, WDiaMapCount, RRSEintersectWDiaMap, SubMapCount, RRSEinSubMap) with *i* in {*set of Questions*} and *j* in {*set of search engines*}.

5.2.2 Experiments

We performed several experiments with the protocol described above. The result of each experiment is represented as a single quintuplet. We selected five questions relating to various diabetes situations that might be encountered by users, and four general search engines, resulting in a total of 20 quintuplet comparisons. We then analyzed the resulting quintuplets to estimate the added value of using DiaMap for information retrieval in the context of diabetes self-management. All experiments were performed between December 2019 and February 2020. The results obtained with search engine queries are likely to change over time.

A diabetes expert selected five different questions reflecting the real-world retrieval of information about diabetes by diabetes patients, health professionals, close relatives and friends. The questions covered several different categories of information highly representative of users' concerns about diabetes. The expert provided the following categories: blogs (prized for self-management), niche-like topics (information about a particular topic), online shopping, hospital information and charity organizations. In accordance with the first step of the protocol, the five categories were expressed as questions, to be used as queries, which were translated into sets of tags as follows:

1) Blogs
   Q_{blog} = blogs for psychological support in gestational diabetes?
   T_{blog} = Blog, psychological support, gestational diabetes.
2) Niche-Like Topics
   Q_{niche} = Is there any association talking about diabulimia?
   T_{niche} = Association, diabulimia
3) Online Shopping
   Q_{com} = Where I can find online shopping for insulin pump decoration?
   T_{com} = E-commerce, company, accessories
4) Hospital Information
   Q_{hos} = Is there a hospital or clinic dedicated to type 1 diabetes?
   T_{hos} = Hospital, clinic, type 1 diabetes, treatment

5) Charity Organizations

Q_{char} = I want to donate some money to diabetes organizations, is there a click-to donate website?

T_{char} = Click-to-donate, charity, advocacy

5.3 Applications on Blogs

Querying DiaMap with T_{blog} provided {W-DiaMap} = {Diabetesramblings.com, Yogafordiabetesblog.com} and the associated submap (see figure 30).

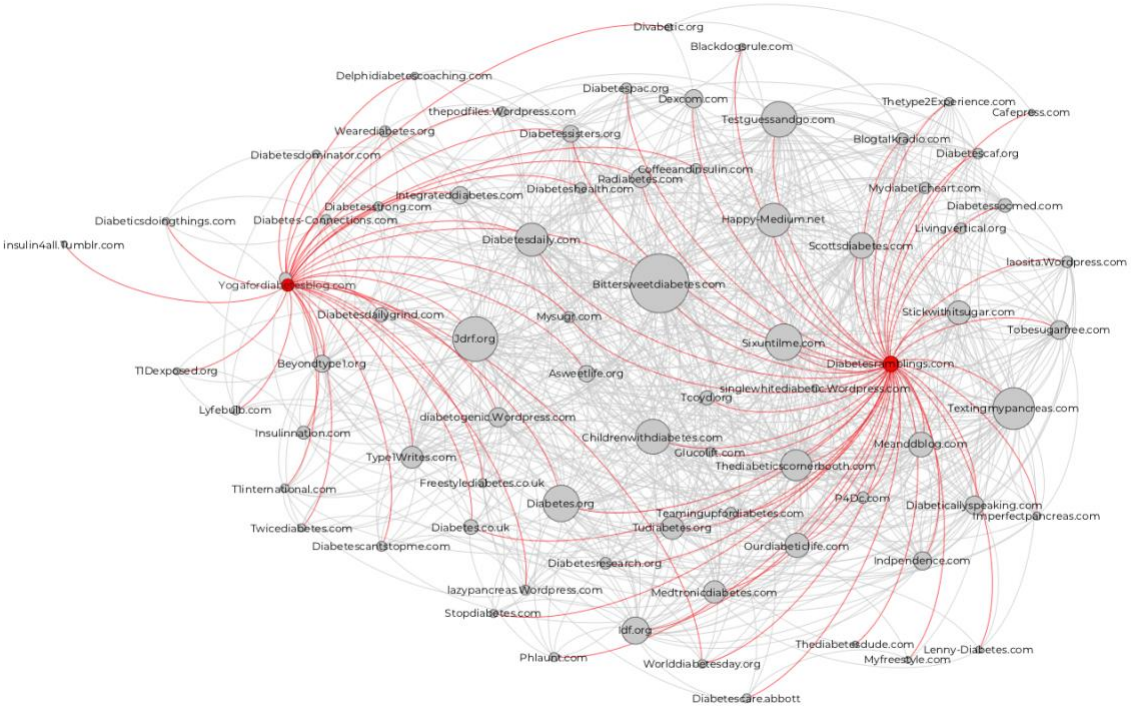


Figure 30 Submap of {W-DiaMap} for T_{blog} with hyperlinked neighbors.

Querying the SEs with Q_{blog} provided the top five websites for each SE, with the associated indicators (Table 7). RRSEs were empty except 2 for Yahoo.

Table 7 Top 5 websites proposed when each search engine was queried with Q_{blog}, and the associated indicators.

Search Engine	TOP 5 Websites	**Relevant**
{W-Google}	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5364143/	No
	https://www.tommys.org/pregnancy-information/pregnancy-complications/gestational-diabetes/gestational-diabetes-and-your-mental-wellbeing	No
	https://www.everydayhealth.com/gestational-diabetes/gestational-diabetes-support-groups.aspx	No
	https://bmjopen.bmj.com/content/8/2/e020462	No
	https://diabetesnsw.com.au/about-diabetes/gestational-diabetes/your-emotional-wellbeing/	No
{W-Bing}	https://www.diabetes.org.uk/About_us/News/Emotional-support	No
	https://www.everydayhealth.com/gestational-diabetes/gestational-diabetes-support-groups.aspx	No
	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3784864/	No
	https://acbrd.org.au/2018/03/29/uk-all-party-parliamentary-group-discussed-emotional-and-psychological-support-of-people-with-diabetes/	No
	https://www.healthline.com/health/diabetes/best-blogs-of-the-year#1	No
{W-Baidu}	https://www.researchgate.net/publication/230574649_Psychological_stress_associated_with_diabetes_during_pregnancy_a_pilot_study	No
	http://xueshu.baidu.com/usercenter/paper/show?paperid=343de48f6f0f8aaa8d9c9cc5c68efffe&site=xueshu_se	No
	http://en.cnki.com.cn/Article_en/CJFDTOTAL-ZDYS201215007.htm	No
	http://xueshu.baidu.com/usercenter/paper/show?paperid=f2d657e9ba8665940b4ea7fd578762c3&site=xueshu_se	No
	https://www.diabetes.co.uk/psychological-support-and-counselling-for-diabetes.html	No
{W-Yahoo}	https://www.everydayfamily.com/blog/just-found-gestational-diabetes/?gdprConsent=1	**Yes**
	https://www.diabetes.org.uk/About_us/News/Emotional-support	No
	https://birthwithoutfearblog.com/2013/06/24/the-truth-about-gestational-diabetes-and-why-its-not-your-fault/	**Yes**
	https://asweetlife.org/psychological-support-the-missing-piece-in-diabetes-care/	No
	https://research.halkidiabetesremedy.org/?hop=3c3q2k11	No

Indeed, although the first five websites listed by Google were all related to gestational diabetes and emotional well-being, none of them were blogs or had any of the properties of this particular web publication format. Indeed, two of the websites were academic articles and the other three were websites of organizations rather than blogs. A similar result was obtained for Bing: none of the first five websites presented had anything to do with blogs. One result similar to those provided by Google, an information webpage about gestational diabetes from everydayhealth.com, was obtained, but, otherwise, no relevant results were obtained. Baidu, proposed four academic articles and one non-academic site, diabetes.co.uk, a community website providing support for

patients with diabetes. Neither of these webpage formats is a blog. The RRSE for Baidu was, therefore, also empty.

Yahoo proposed two websites accurately responding to the question. One was the blog of everydayfamily.com and the other was the gestational diabetes webpage from birthwithoutfearblog.com. Both also deal with pregnancy, birth, and the postpartum period. Thus, RRSEcount was 2 for Yahoo.

We found no {W-DiaMap} among the first five webpages proposed by any of the four search engines for the Q_{blog} query. Hence, RRSEintersectWDiaMap = 0 for all the search engines. SubMapCount = 76 for this submap, only one website (asweetlife.org) from {W-Yahoo} was present in the submap, so RRSEinSubMap = 1 for Yahoo but 0 for the other three search engines.

5.4 Applications on Niche-Like Topic

Querying DiaMap with T_{niche} provided {W-DiaMap} = {Diabulimiahelpline.org, Dwed.org.uk, Wearediabetes.org}, and the associated submap (see figure 31).

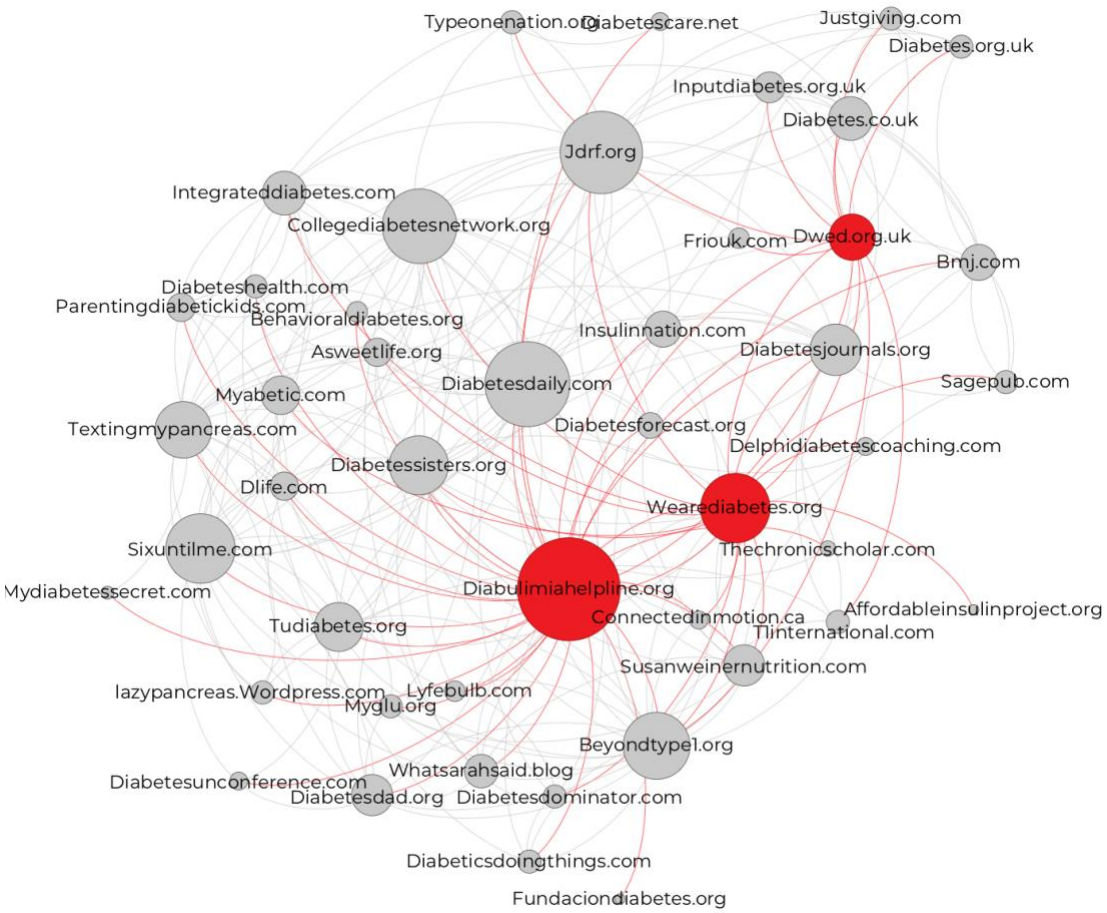


Figure 31 Submap of {W-DiaMap} for T_{niche} with hyperlinked neighbors.

Querying the SEs with Q_{niche} provided the top five websites for each SE with the associated indicators (Table 8). RRSEcount was 4 for Google, 1 for Bing, 0 for Baidu and 3 for Yahoo.

Table 8 Top 5 websites proposed when each search engine was queried with Q_{niche}, and the associated indicators.

Search Engines	TOP 5 Websites	**Relevant**
{W-Google}	https://www.nationaleatingdisorders.org/diabulimia-5	**Yes**
	https://www.diabetes.org.uk/guide-to-diabetes/life-with-diabetes/diabulimia	**Yes**
	https://www.diabetes.org.uk/guide-to-diabetes/life-with-diabetes/diabulimia/signs-of-diabulimia	**Yes**
	https://psych2go.net/diabulimia-the-eating-disorder-no-one-talks-about/	No
	http://www.diabulimiahelpline.org/meet-dbh.html	**Yes**
{W-Bing}	https://www.diabetes.org.uk/Guide-to-diabetes/Life-with-diabetes/Diabulimia	**Yes**
	https://www.webmd.com/diabetes/what-is-diabulimia#1	No
	https://www.youtube.com/watch?v=PrdLOdEqMxU	No
	https://beyondtype1.org/the-truth-about-diabulimia/	No
	https://www.mic.com/articles/136237/diabulimia-eating-disorder-risks-causes-myths-and-facts	No
{W-Baidu}	http://xueshu.baidu.com/usercenter/paper/show?paperid=f705f49a94d3357bb0abea8b8eaa82b5&site=xueshu_se	No
	http://xueshu.baidu.com/usercenter/paper/show?paperid=9628d5d6380c39409f2e0c06e9962bd3&site=xueshu_se	No
	https://www.researchgate.net/publication/236866468_Full_moon_days_and_crime_Is_there_any_association	No
	http://xueshu.baidu.com/usercenter/paper/show?paperid=2de3095a18a6a8c76bfd7ae400fc3d61&site=xueshu_se	No
	https://www.researchgate.net/publication/308578457_Folate_and_Cancer_Is_There_Any_Association	No
{W-Yahoo}	https://www.therecoveryvillage.com/mental-health/diabulimia/	**Yes**
	https://www.adwdiabetes.com/articles/diabulimia-type-1-diabetes-eating-disorder	No
	https://www.webmd.com/diabetes/what-is-diabulimia#1	No
	https://www.diabetes.org.uk/Guide-to-diabetes/Life-with-diabetes/Diabulimia/Signs-of-Diabulimia	**Yes**
	https://www.nationaleatingdisorders.org/sites/default/files/ResourceHandouts/Diabulimia.pdf	**Yes**

All the search engines performed relatively well for this question. We found three websites relating to this question among the top five websites proposed by Google: diabulimiahelpline.org, nationaleatingdisorders.org and diabetes.org.uk, which do no focus specifically on diabulimia but nevertheless provide guidance for people with this food disorder. RRSEcount = 4 for Google, the highest value obtained in this study.

The first website proposed by Bing was related to the question, whereas the other four were not relevant. So RRSEcount = 1 for Bing for this question. Unsurprisingly, with Baidu, the top five websites were all academic articles, providing some insight into how Baidu built its corpus for English resources. However, none of these articles had anything to do with diabulimia, so

RRSEcount = 0. For Yahoo, three of the five websites proposed are related to the question. The first corresponded to the recovery village, which is a treatment center rather than an association. Another two associations were identified: diabetes.org.uk and nationaleatingdisorders.org, the same two organizations proposed by Google. Therefore, RRSEcount = 3 for Yahoo, for this question.

For our next indicator, RRSEintersectWDiaMap, only one website was found in both {W-Google} and {W-DiaMap} for Q_{niche}, so RRSEintersectWDiaMap = 1 for Google. For the other search engines, no website was found in both RRSE and {W-DiaMap}, so RRSEintersectWDiaMap = 0.

SubMapCount = 46 for this submap, indicating a less populated immediate neighborhood compare the blogs' submap. This may be because blogs frequently hyperlink widely to other resources, whereas the associations dominating this submap are more careful or selective about the resources they link to, resulting in higher overall relevance. This is supported by the results obtained: RRSEinSubMapCount = 3 for Google, RRSEinSubMapCount = 1 for Bing and RRSEinSubMapCount = 1 for Yahoo, which are high relative to the values obtained for the other questions.

5.5 Applications on Online Shopping

Querying DiaMap with T_{com} provided {W-DiaMap} = {cafepress.com, society6.com} and the associated submap (see figure 32).

Figure 32 Submap of {W-DiaMap} for T_{com} with hyperlinked neighbors.

Querying the SEs with Q_{com} provided the top five websites for each SE with the associated indicators (Table 9). RRSEcount was 3 for Google, 0 for Bing, 0 for Baidu and 2 for Yahoo.

Table 9 Top 5 websites proposed when each search engine was queried with Q_{com}, and the associated indicators.

Search Engines	TOP 5 Websites	**Relevant**
{W-Google}	https://www.alibaba.com/premium/insulin_pump.html	No
	https://www.medtronicdiabetes.com/products/insulin-pump-style-and-accessories	**Yes**
	https://pumppeelz.com	**Yes**
	http://shop.pepmeup.org/product/medtronic-minimed-tropical-640g-cover-skin	**Yes**
	https://www.pinterest.com/pin/456130268505071645/	No
{W-Bing}	https://www.adwdiabetes.com/category/insulin-pump-supplies	No
	https://asweetlife.org/shopping-for-a-new-insulin-pump-whos-selling/	No
	https://www.diabetesforum.com/insulin-pumps/74354-shopping-pump.html	No
	https://www.totaldiabetessupply.com/collections/insulin-pump-supplies	No
	https://www.pinterest.com/explore/insulin-pump/	No
{W-Baidu}	https://www.alibaba.com/showroom/home-certain.html	No
	ps://www.researchgate.net/publication/303867676_Risk_and_Protective_Factors_for_Childhood_Asthma_What_Is_the_Evidence	No
	http://www.nosdiet.com/	No
	https://www.globalsources.com/factory/Retractable.html	No
	https://fanyi.baidu.com/?aldtype=23#en/zh/Where%20I%20can%20find%20online%20shopping%20for%20insulin%20pump%20decoration	No
{W-Yahoo}	https://www.redbubble.com/shop/insulin+pump+stickers	**Yes**
	https://www.adwdiabetes.com/category/insulin-pump-supplies	No
	https://shop.mybluprint.com/sewing/project/insulin-pump-pouch/33686	No
	https://www.medtronicdiabetes.com/products/insulin-pump-style-and-accessories	**Yes**
	https://www.totaldiabetessupply.com/collections/insulin-pump-supplies	No

The top five results for Google included three websites relating to this question, so it would be easy for users to find decoration for insulin pumps with this search engines. The other two websites were not related to diabetes: the commercial website Alibaba and the Pinterest social community. RRSEcount = 3 for Google. Bing and Baidu results were entirely unsatisfactory, with no interesting websites related to the question identified. However, rather than just academic

articles, Baidu, in this case, proposed general online commercial platforms, such as Alibaba, or globalsources.com. RRSEcount was zero for these sites. Yahoo provided good results, identifying two relevant websites. The other three non-related websites were websites selling insulin pump supplies rather than specifically decoration for insulin pumps. RRSEcount = 2 for Yahoo.

There was no overlap between {W-DiaMap} and RRSE, so RRSEintersectWDiaMap = 0 for the 4 all four search engines.

SubMapCount = 2 for this query, because there was no other website linked to it in the DiaMap. This is often the case for commercial websites, which use a strategy of encouraging users to remain rather than promoting other websites. This behavior is often criticized in online communities.

As a result of the restrained SubMapCount, no RRSEinSubMap was zero.

5.6 Applications on Hospital Information

Querying DiaMap with T_{hos} provided {W-DiaMap} = {Barbaradaviscenter.org, Clinidiabet.com, Hopkinsmedicine.org, Massgeneral.org, Rch.org.au} and the associated submap (see figure 33).

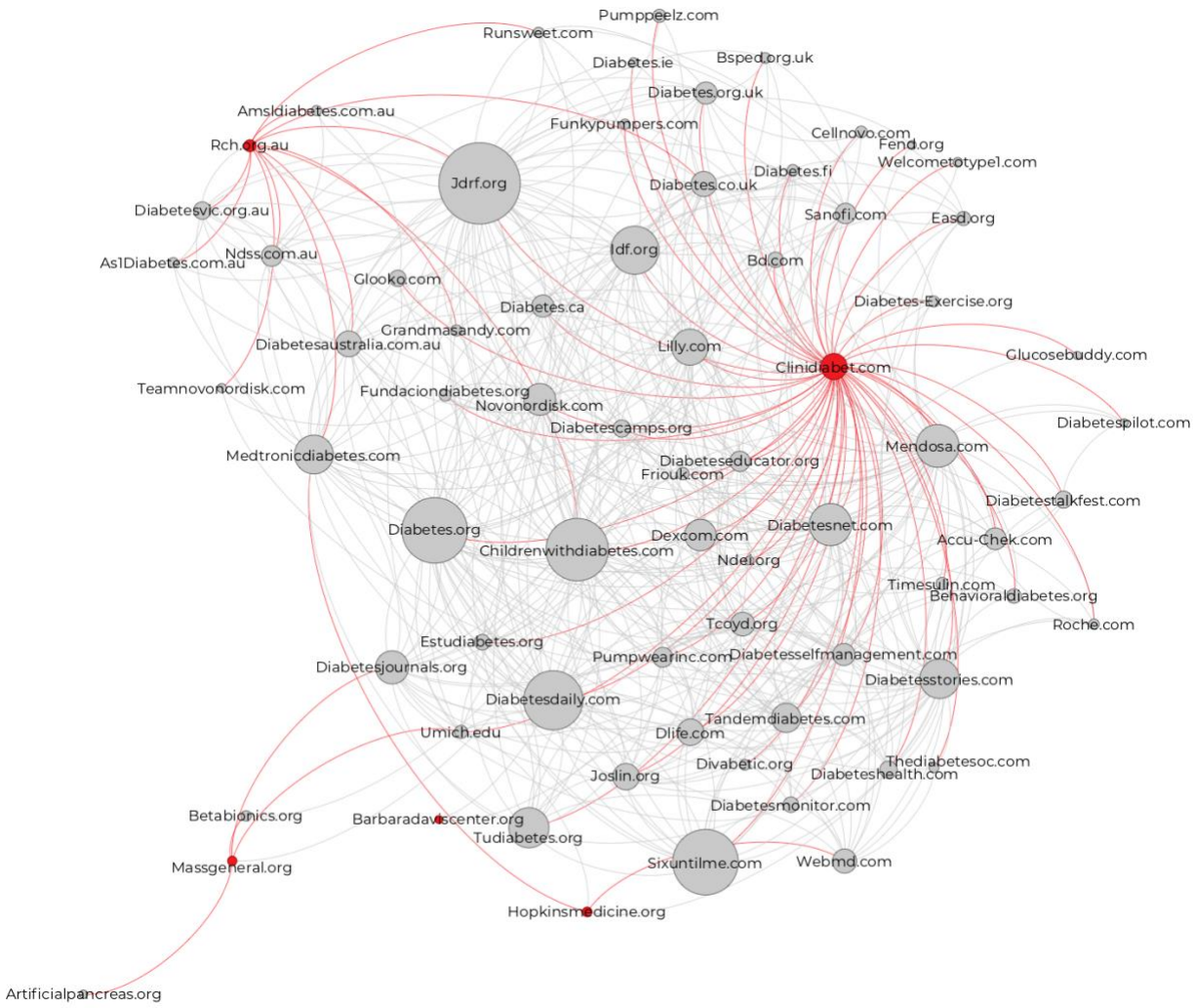


Figure 33 Submap of {W-DiaMap} for T_{hos} with hyperlinked neighbors.

Querying the SEs with Q_{hos} provided the top five websites for each SE with associated indicators (Table 10). RRSEcount was 2 for Google, 1 for Bing, 0 for Baidu and 2 for Yahoo.

Table 10 Top 5 websites proposed when each search engine was queried with Q_{hos}, and the associated indicators.

Search Engines	TOP 5 Websites	**Relevant**
{W-Google}	https://www.armi.org.au/	No
	https://www.idf.org	No
	https://www.em-consulte.com/en/article/276237	No
	https://www.lewishamandgreenwich.nhs.uk/diabetes-services-in-lewisham/	**Yes**
	https://www.uclh.nhs.uk/OurServices/ServiceA-Z/CYPS/PDIAB/Pages/Home.aspx	**Yes**
{W-Bing}	https://forum.diabetes.org.uk/boards/threads/best-clinic-hospital-for-type-1s-in-london.64676/	No
	https://beyondtype1.org/clinical-trials-and-the-type-1-diabetes-cure/	No
	https://jdrf.org.uk/information-support/living-with-type-1-diabetes/healthcare-support/your-healthcare-team/	No
	https://news.sanfordhealth.org/news/clinical-trial-milestone/	No
	https://www.rch.org.au/diabetes/type-1-diabetes/	**Yes**
{W-Baidu}	http://www.doc88.com/p-6962505741376.html	No
	http://xueshu.baidu.com/usercenter/paper/show?paperid=d323e90d8e471079f67f050489da5ea7&site=xueshu_se	No
	https://max.book118.com/html/2017/0516/107016794.shtm	No
	https://www.doc88.com/p-0179782302570.html	No
	https://www.researchgate.net/publication/273191266_Do_Healthcare_Workers_Adhere_to_Diabetes_Clinical_Care_Guidelines_A_Study_at_a_National_Hospital_Kenya	No
{W-Yahoo}	http://www.healthtalk.org/young-peoples-experiences/diabetes-type-1/what-happens-diabetes-clinic	No
	https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/diagnosis-treatment/drc-20353017	**Yes**
	https://www.uabmedicine.org/diabetes	No
	https://beyondtype1.org/clinical-trials-and-the-type-1-diabetes-cure/	No
	https://my.clevelandclinic.org/health/diseases/17666-type-1-diabetes-in-children	**Yes**

The top five websites from Google included two positive answers relevant to the question. Lewishamandgreenwich.nhs.uk and uclh.nhs.uk are both hospitals and mention on their websites that they care for patients with type 1 diabetes. RRSEcount = 2 for Google. The one positive

website suggested by Bing was rch.org.au. This is the official website of the Royal Children's Hospital in Melbourne, which improves the health and wellbeing of children and adolescents through leadership in healthcare, research and education. This institution also explicitly mentions diabetes on its website so RRSEcount = 1 for Bing. Baidu found nothing relating to the question; RRSEcount = 0. Finally, Yahoo provided two positive answers: Mayoclinic.org is the Mayo Clinic Health System in America and clevelandclinic.org is a nonprofit multispecialty academic medical center. RRSEcount = 2 for Yahoo, for this question.

Only one website was common to RRSE-Bing and {W-DiaMap} for Q_{hos}, so RRSEintersectWDiaMap = 1 for Bing. For the other search engines, no website was common to both RRSE and {W-DiaMap}

The submap is quite substantial, with SubMapCount = 71. Most of the websites are neighbors of rch.org.au and clinidiabet.com, and these two hospitals seem to have an online strategy designed to make them easy to find online, with referencing by many organizations and associations.

Finally, RRSEinSubMap = 1 for Bing. Other websites from the search engine SERPs were found in the submap but did not belong to RRSE and had ranks among the results. For example, Hopkinsmedicine.org was found on the fifth page of results for Yahoo. Similarly, Rch.org.au was found on the third pages of results for both Google and Yahoo. RRSEinSubMap was zero for Google, Baidu and Yahoo.

5.7 Applications on Charity Organizations

Querying DiaMap with T_{char} provided {W-DiaMap} = {Justgiving.com, Everydayhero.com, Jdrf.org} and the associated submap (see figure 34).

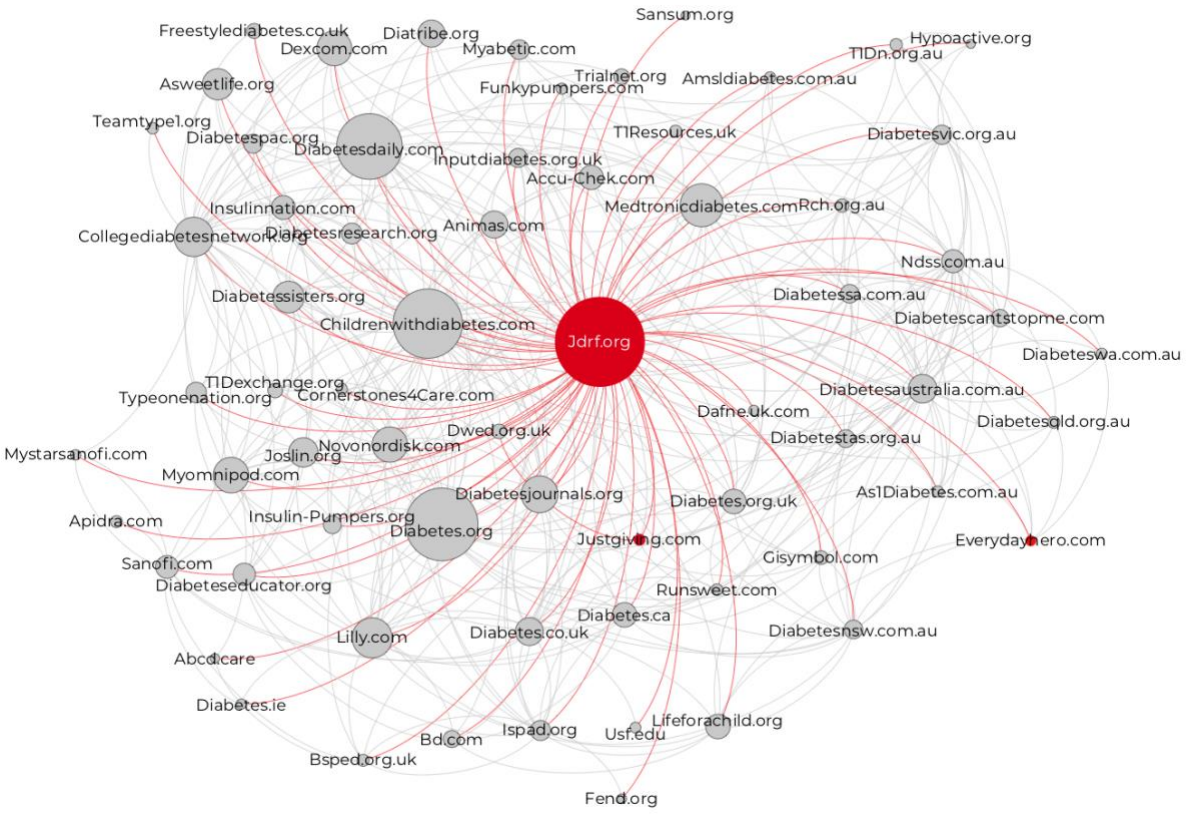


Figure 34 Submap of {W-DiaMap} for T_{char} with hyperlinked neighbors.

Querying the SEs with Q_{char} provided the top five websites for each SE, with associated indicators (Table 11). RRSEcount was 1 for Google, 2 for Bing, 0 for Baidu and 1 for Yahoo.

Table 11 Top 5 websites proposed when each search engine was queried with Q_{char}, and the associated indicators.

Search Engines	TOP 5 Websites	**Relevant**
{W-Google}	https://www.michaeljfox.org/donate/our-goal-urgently-needed-cure?smcid=ap-a1b1R0000086fJ2&	No
	https://www.moco2makuyu.org	No
	https://www.cantransplant.ca/organ	No
	https://www.populationmedia.org/donate-now/	No
	https://www.thenonprofits.com	**Yes**
{W-Bing}	https://leonormarchand.wixsite.com/runfordiabetes/copie-de-donate#!	**Yes**
	https://donate.diabetes.org.uk/donate/~my-donation	**Yes**
	https://dollarsanity.com/donate-without-spending/	No
	https://www.consumerreports.org/charities/best-charities-for-your-donations/	No
	https://www.move.org/how-to-donate-your-used-items/	No
{W-Baidu}	This query yielded no result	No
{W-Yahoo}	http://www.diabetesforecast.org/2018/06-nov-dec/how-to-donate-unused-diabetes.html	No
	https://asweetlife.org/giving-to-diabetes-charities-where-does-the-money-go/	No
	https://www.consumerreports.org/charities/best-charities-for-your-donations/	No
	https://diatribe.org/do-you-have-extra-diabetes-supplies-you-no-longer-need	**Yes**
	https://www.webmd.com/a-to-z-guides/kidney-donation-steps#1	No

The top five websites from Google included only one relevant website. Thenonprofits.com is basically a donation website and the specific subdomain for donation related to diabetes research is just one of many. RRSEcount = 1 for Google. We obtained two positive websites with Bing: leonormarchand.wixsite.com and diabetes.org.uk. The first is simply a click-to-donate site for diabetes research to help cure type 1 diabetes and the second is the part of the British Diabetic Association website dealing with donation. RRSEcount = 2 for Bing. Baidu generated no results at all for this question. For Yahoo, only one website relating to this question was identified: a website for online donations for all type of diabetes. RRSEcount = 1 for Yahoo.

No website was common to both RRSE and {W-DiaMap} for Q_{char}, given RRSEintersectWDiaMap = 0 for Q_{char}.

SubMapCount = 67 for this question, with jdrf.org a highly popular NGO involved in charity work. This popularity is highly visible in the submap. Nevertheless, RRSEinSubMap = 1 only for Yahoo. Jdrf.org was proposed by the other search engines, but lower down the first page for Google, on the fifth page for Bing and fourth page for Yahoo. It was not only the top five webpages of any search engine queried with Q_{char}.

5.8 Global comparison with Search Engines and DiaMap

Results for the 5 questions x 4 search engines = 20 experiments are reported according to the quintuplet representation described in the protocol (Table 12). These results will be discussed in Chapter 6.

Table 12 Results for the 20 experiments, presented according to the quintuplet presentation described in the protocol.

		Q1 *Blogs*	**Q2** *Niche Topics*	**Q3** *Online Shopping*	**Q4** *Hospital Information*	**Q5** *Charity Organization*
RRSEcount	Google	0	4	3	2	1
	Bing	0	1	0	1	2
	Baidu	0	0	0	0	0
	yahoo!	2	3	2	2	1
WDiaMapCount		2	3	2	5	3
RRSEintersectWDiaMap	Google	0	1	0	0	0
	Bing	0	0	0	0	0
	Baidu	0	0	0	0	0
	yahoo!	0	0	0	1	0
SubMapCount		76	46	2	71	67
RRSEinSubMap	Google	0	3	0	0	0
	Bing	0	1	0	1	0
	Baidu	0	0	0	0	0
	yahoo!	1	1	0	0	1

1) The higher is RRSEcount, the better is the search engine at proposing relevant websites. For instance, if RRSEcount = 3, it means that the first five websites proposed by the SE include three relevant for Q.
2) WDiaMapCount is the cardinality of {W-DiaMap}, identifying relevant websites from the DiaMap.
3) RRSEintersectWDiaMap is the cardinality of RRSE ∩ {W-DiaMap}. It shows results common to both corpuses.
4) SubMapCount counts the websites in the submap. The higher SubMapCount, the larger the neighborhood of the relevant website of the DiaMap.
5) RRSEinSubMap is the intersection of RRSE and the submap. If RRSEinSubMap is larger than RRSEintersectWDiaMap, then the submap proposes relevant results not fully tagged as such in DiaMap. This shows that the submap can yield relevant results even if the fully tagged search is not complete.

6 Chapter 6 Discussion and Conclusion

In this chapter, we will combine the previous chapters to discuss about our methods, results and limitations of this manuscript. We will also talk about the perspectives and the future work.

6.1 Discussion about methods

In Chapter 3, when we presented the methodology how to visualize the map of hyperlinks structure on diabetes communities, we used two state-of-arts tools Hyphe and Gephi. The proposed methodology follows the 6 steps based on several decision criteria and we explain some other possibilities here one by one.

First, gathering a list of starting websites to feed Hyphe. As the objective is to explore all aspects of the community, we selected a variety of sources starting with diabetes experts, search engines and diabetes community social network pages suggestions to feed Hyphe. Indeed, to explore an online community consisting in topic related websites, we need to refine entry points into the community. However, there are still some alternatives that can try to collect the starting websites and as we started with English solo language websites first, that leads to a majority of English-domain websites in the final map despite having some multilingual websites coming from US (using English and Spanish) or Canada (using English and French). Changing the different language websites as initial websites are very attempting to describe the whole map of diabetes in the digital world. Actually, to try with Chinese websites will be interesting due to the poor performances from Baidu search engine during the comparison with DiaMap in Chapter 5.

Second, choice of depth and the criteria to stop crawling. In this work, we decided to consider a crawl at depth 1 to avoid too much noise. Due to the crawler limitation and the relatively small number of websites chosen to be presented in the final map, it cannot show the whole world with diabetes but just part of it. As the web network follows a power law degree distribution, a crawl at depth 3 can potentially gather thousands of websites making it overwhelming for a human to properly assess each membership to the targeted diabetes community. However, a number of website overwhelming for a human being can be assessed using machine learning. Once we can

use Nature Language Processing (NLP) to extract accurate and useful tags to annotate each website automatically, a deeper crawl would undoubtedly have more information covered and it would be the ideal way to expend the size of DiaMap in the future.

Third, cleaning process to feed each iteration with an IN database. We classified the contents of the web entities strictly related to diabetes as IN and when nothing indicates a potential diabetes related content, we classified it as OUT. UNDECIDED websites are those ambiguous websites that mention diabetes among other topics. In this study, we filter the database to only retain the IN websites. Actually, to keep crawling UNDECIDED websites can help us distinguish the neighbor topic of diabetes like nutrition websites in the digital world. Which contents are close to diabetes or which industry is combined with diabetes? Then we can easily figure out the frontier of the map with diabetes online communities.

Fourth, choice of the final pattern in Gephi and applying the community detection algorithm. After we imported the network into Gephi, we used the layout "Force Atlas 2" to adapt the scale free networks to spatialize the nodes and their edges. The proximity of the nodes is the result of the force vector based layout algorithm only and do not demonstrate any physical proximity whatsoever. To provide a first view at the sub-communities inside the diabetes community, we applied a community detection algorithm aimed at detecting clusters of similarly connected nodes. It does so by maximizing the quality metric known as modularity over all possible partitions of a network. Modularity measures the difference between the edge density in the partition and a randomized graph with the same number of nodes and the same degree distribution. In another word, if we use a different layout or if we change choose to apply a different classification algorithm, would we have different final 5 classes result?

Fifth, semantic approach based on tags. In this manuscript, we used a semantic approach based on tags chosen by one diabetes expert. After the annotation, we found some tags are too specific and unnecessarily complicate the annotation process. For example, tags as clinic and hospital can be replaced by healthcare facilities to just simplify the tags. They are maybe useful from a user point of view but are not good enough to become the discriminating mechanism to identify regions or localities inside the corpus. In Chapter 4, we mentioned there are three main

approaches to retrieve, extract and describe the semantic content of websites, terminology, ontology and Nature Language Processing. We saw some potential in ontology but since a modular ontology of the stakeholders doesn't yet exist in the context of the online diabetes field, and building such ontology is very time consuming in itself, we adopted a terminology approach. In the future, if we can complete ontology of the stakeholders of diabetes, semantic approach can be more precise and accurate.

6.2 Discussion about results

This manuscript demonstrates the potential approach in analyzing and visualizing the diabetes-related websites network.

Using Hyphe to extract 430 specific and relative websites, people with diabetes can easily navigate most relevant websites to obtain information and knowledge about their conditions or different aspects of it.

Furthermore, such a strong community can offer support from a knowledge and a psychology point of view. Search engine like Google can find global websites by considering their authority in the network such as general organizations like International Diabetes Federation (IDF), American Diabetes Association (ADA), Diabetes Australia, etc. But they cannot offer people with diabetes detailed information to help people manage their diabetes and improve the quality of their lives efficiently. 80% of the iceberg is still under the sea and is difficult to reveal. Hyphe and our methodology is proven reliable to explore the localities, and we were able to find more interesting websites such as personal blogs sharing the stories of the real diabetes daily lives including their struggles to combat diabetes. We can also find different diabetes online communities focusing on the different aspects, nutrition, sports, camping, etc. In addition, some websites are more difficult to detect when they talk about some rare issues, such as diabulimia which is an eating disorder related to type 1 diabetes.

We focused on Social Network Analysis (SNA) as new methodological tool to study web-based topic related network. We provided some criteria for collecting hyperlinked data. SNA is an extension of traditional network analysis because it focuses on the structure of a social system based on the relationship between actors. The difference between hyperlinked network, social network and traditional network analysis is the use of hyperlink in webpages as the trace of a relationship whose intentions remain hidden. We used data obtained from websites. This data is the result of the analysis of the content of HTML (Hyper Text Markup Language) data that is subject to improvement. Many websites nowadays hide their link under a layer of interaction and build the link with a computer script when users click on them rather than plainly using the anchor tag. There are two consequences here: one is that our network is dependent of techniques that were not designed for it specifically and web network indicators might be a future source of investigation if not innovation; second is that the edges of our network that materialize the relations might also change in the future, to include more of the original intention if semantic web gets more popular or to be harder to build with large scale harvesting if HTML technics shift toward hiding links. Both consequences would affect edge construction and require further work to include it in the network modification.

Semantic interpretation of the map with diabetes-related websites assumes that all stakeholders or actors involved in diabetes have a similar type of relation or connections. This can reveal the web as an organized world of communities (including the diabetes community) rather than a randomly organized network. Moreover web network specialists argued that social network analysis not only reveals the social structure of the Internet, but can also be used to examine the communication strategy and pattern among actors (Park & Thelwall, 2003).

As we mentioned in the Chapter 1, in order to address the relationship with diabetes stakeholders, we examined how traditional studies from the network explorations can be used in the context of diabetes. We found that, in the class 1, it mainly includes the healthcare system and associations (society and federation) but rarely includes the blogs of people with diabetes. That means the modularity between the healthcare system, associations and individuals are less close than others. It somehow shows the healthcare system, associations are still lacking the patients' voice. This finding is consistent with the phenomenon that the voice from patients is usually easily

ignored (Skovlund & Peyrot, 2005). Our work supports the need for future research on how to merge the patients' opinions to the healthcare professionals.

Interestingly, in the class 3, it mainly includes blogs which focus on self-management and psychological support with absolutely no healthcare system inside. That reveals blogs focus on immediate management issues than talking about treatments or the technology devices (Oser et al., 2017), but to our knowledge, so far there is only study on the analysis of caregivers for type 1 diabetes. In our work, we covered 129 blogs which contains type 1, type 2 diabetes' and also caregivers'. People with diabetes are easily to release their emotion and get support from peer supports.

Results show a low prediction performance by using tags to provide a semantic explanation of the clusters of diabetes-related websites obtained in our work. While looking at the tags distribution, this result reflects the fact that some tags are specific to cluster 1 or 2 and they are the ones with maximum weight found from 7 methods. This means that two of our clusters do have an identity that is relatively easy to predict with tags. However, even for cluster 1 or 2, it is hard to explain the prediction by a specific combination of the tags (a semantic signature). This leads us to two hypotheses: either we did not select the proper tags to explain the classes, either our dataset is not big enough to figure out the semantic meaning behind it.

As we discussed the semantic approach based on tags before, one perspective will be to use alternative approaches, like Natural Language Processing (NLP), to automatically extract the most relevant tags from the whole set of websites to improve the prediction rate. Another approach could be to learn the most appropriate tags from the set of websites for each cluster.

In addition, we found that the main two clusters with the best prediction rate are also containing the largest number of websites. Indeed, the resolution limitation of modularity makes the algorithms unable to detect small communities (Lancichinetti & Fortunato, 2011), even for one of the best model in our study which is Random Forest. This leads us to a new hypothesis that the current network is not big enough for machine learning to predict the right cluster. In future work, using the crawling procedure, we intend to extend the number of DiaMap to above 5000+ diabetes-related websites, with the objective to predict more accurately the related clusters. It will also help us to get a more appropriate size for each cluster for learning methods.

However, this will raise the big issue of time consuming for the manual annotation process. Indeed, considering 430 websites took almost 1 month to tag so that tagging 5000+ websites seems to be a too time-consuming job. One solution to this is to again, use NLP solutions for the automatization of the annotation process. If we can use NLP to get accurate tags, can we also use it to facilitate the annotation process? Somehow, the most difficult part for teaching machine to do the annotation is not only to detect the presence of a tag in the web page but rather to understand the meaning behind. When experts annotate manually, usually the first step is to read "About Us" or "Who we are" to get the general idea about the website. For example, like www.jdrf.org, in "About Us" part, it is clearly written "JDRF is the leading global organization funding type 1 diabetes (T1D) research." And their mission is "Improving lives today and tomorrow by accelerating life-changing breakthroughs to cure, prevent and treat T1D and its complications." *(https://www.jdrf.org/about. Archived at: http://www.webcitation.org/77PNGOZxs)* So it is easier to find the tags for Non-profit, Type1, Association, Prevention, Treatment and Complications. But with some other websites, it takes time to read the webpages to get the main information to annotate. Also with the different ways to express the same meaning, like type1 diabetes, T1D, insulin dependent diabetes, Juvenile diabetes, we need to collect the enough words in our corpus to describe the contents.

About accessing DiaMap for information retrieval and domain awareness, we used in Chapter 5, metrics to compare current search engines and DiaMap, Google and Yahoo had similar, good performances, with a mean RRSEcount of 2 for the five questions. The third-ranked SE in terms of performance was Bing with a mean RRSEcount of 0.8 for the five questions. Baidu proposed no relevant results, but this is not particularly surprising because it focuses on the Chinese search engine market.

WDiaMapCount consistently displayed more relevant results than RRSEcount, with a mean of three relevant results per question. This value is higher than for the best search engines. DiaMap therefore provides users with more relevant results than SEs. Furthermore, as DiaMap does not display unrelated websites, there is no mixed set of relevant and non-relevant websites to choose from. This improves accuracy and decreases any potential loss of time.

RRSEintersectWDiaMap does not appear useful at first glance, because there is almost no intersection between the two sets for a question translated fully into tags (all tags have to be true). For the five queries considered, DiaMap missed only two good websites proposed by search engines. It had hits for almost all the relevant websites and displayed clear matches to questions. One explanation for this is that general search engine algorithms do not rank results according to how well the website semantically answers the question. Instead, they use a wide range of network topology indicators and user behaviors. The semantic match is just one of a number of indicators. This is partly due to the size of the corpus, which is enormous. Although advantageous in some circumstances, this huge size is difficult to handle when posing a question related to a specific scenario. The reason is the most popular and best-known websites (usually with an excellent network and social strategy) tend to be promoted, regardless of the potential underlying communities. DiaMap proposes an alternative approach based on mapping of the diabetes community, with this community used as a corpus. Website popularity is thus a secondary criterion after the semantic match. These differences account for the lack of intersection between the two sets in most cases.

SubMapCount was also consistently high (46 to 76), even starting with two to five websites by excluding commercial websites. Commercial websites are not generally well-integrated into communities and do not, therefore, have a large submap. DiaMap does not focus on commercial sites, potentially accounting for the scarcity of relevant websites for this question. These results also clearly identify ways of improving DiaMap by extending it while carefully checking which websites are added.

RRSEinSubMap was determined to demonstrate that DiaMap also covered the websites proposed by search engines. We used the network properties of DiaMap to retrieve five submaps; each submap was generated by extracting all the neighbors (i.e. sharing a hyperlink) of each {W-DiaMap} result. This provided information about the immediate proximity of the websites potentially encountered by users during navigation. The first five websites in RRSE had a larger presence in the submap than in {W-DiaMap} as RRSEinSubMap > RRSEintersectWDiaMap. Thus, DiaMap covers the websites proposed by search engines, but probably with different tags.

DiaMap identified 15 relevant websites in total with five queries, whereas the numbers of RRSE in the submaps were three for Google, two for Bing, zero for Baidu and three for Yahoo. Of course, it would not be fair to say that none of the other websites identified by Google were relevant. Indeed, with the third question, search engines gave better results than DiaMap. However, the core and accurate websites provided by DiaMap genuinely addressed the questions posed, and it takes time for users to go through SE results in detail to find the correct information. In search engine-driven searches, 88% of users preferentially click on the first three results, regardless of screen size and what they were looking for (Kim et al., 2015). Search engines can propose a number of websites in response to each query. However, it requires considerable time and knowledge for users to read each website and determine its reliability and relevance for themselves. DiaMap is more like curated content from diabetes online communities. It helps people to find more diabetes-focused websites more accurately, using tags and classes to distinguish between them. DiaMap can thus be seen as a specialist search engine focus on all aspects of diabetes online, to help people with their diabetes.

Given the superior quality of results and precision of DiaMap for diabetes questions, we plan to improve it further, by incorporating the use of natural language into queries instead of the set of tags, and automatically annotating the content of websites with natural language processing (NLP). It may also be useful to add other languages, such as Chinese, to provide accurate results on diabetes to a larger audience.

Search engines have emerged as a major force in the information economy in recent decades. They have significant power to shape the behavior and perceptions of users. Search engines affirmatively control their users' experiences, which has the consequence of skewing search results (a phenomenon called "search engine bias") (A De Corniere, G. Taylor, 2014). We propose an alternative for those searching for useful information of diabetes online and provide evidence to fuel the debate on "search engine bias".

Finally, search engines display results in SERPs, which are carefully designed for a general audience. However, SERPs are prone to frequent changes, particularly if the interpersonal variability of results that personalization promotes is taken into account. We see an opportunity

for an authoritative display of results for health-related information retrieval providing a real alternative for stable navigation. DiaMap is a map-like visualization of network resources. As such, it does not change much over time and provides access to all the available information at a glance. The best way to update and maintain the map with the latest relevant and active websites is the main issue here. In the future, we plan to add a monitoring process to ensure the retention of active websites, addition of new relevant websites and removal of inactive websites.

6.3 Conclusion

We have successfully proved that all stakeholders involved in diabetes have connections on the World Wide Web as an organized world of communities rather than being a randomly organized network. The result showed that diabetes online community has its own space and is organized with clusters and distribution of websites of relative importance. We also applied a community detection algorithm aimed at detecting clusters of similarly connected nodes to provide a first view at the sub-communities inside the diabetes community. In the end, we got DiaMap associated with 5 discovered clusters.

DiaMap presents a map-like visualization of information concerning diabetes-related websites, providing a picture of diabetes in the digital world. DiaMap differs from traditional search engines by presenting the whole picture and using tags to identify relevant websites. It changes the way in which online diabetes information is navigated and provides an alternative to general search engines for obtaining domain-specific information of various categories and on various media, to match the user's expectations.

In this work, we provided a replicable practical methodology which combines crawling and visualization tools to analyze a web-based network of diabetes. It is a novel promising way to analyze further chronic disease stakeholders' organizations and communities networking relationships and provide a new way to find relevant information on a specific topic. However, it is just the entry point to present how to combine the network visualization and both topological and semantic community detection to do the network analysis domain in diabetes. This new

approach can be extended to network analysis for other chronic diseases helping patients and their families to navigate in a traditionally search engine domain digital world.

Reference

About diabetes. World Health Organization. Archived from the original on 31 March 2014. Retrieved 4 April 2014.

A Coefficient of Agreement for Nominal Scales—Jacob Cohen, 1960. (n.d.). Retrieved March 24, 2020, from https://journals.sagepub.com/doi/abs/10.1177/001316446002000104?journalCode=epma

A Comparison of Glyburide and Insulin in Women with Gestational Diabetes Mellitus | NEJM. (n.d.). Retrieved September 25, 2019, from https://www.nejm.org/doi/full/10.1056/nejm200010193431601

Adamic, L. A., & Adar, E. (2003). Friends and neighbors on the Web. Social Networks, 25(3), 211–230. https://doi.org/10.1016/S0378-8733(03)00009-1

Agreement and Information in the Reliability of Coding: Communication Methods and Measures: Vol 5, No 2. (n.d.). Retrieved April 5, 2019, from https://www.tandfonline.com/doi/abs/10.1080/19312458.2011.568376

Al. (eds, M. F. E., Chen, L., Friedman, C., Chen, L., & Friedman, C. (2004). Extracting Phenotypic Information from the Literature via Natural Language Processing.

Al Omran, F. N. A., & Treude, C. (2017). Choosing an NLP Library for Analyzing Software Documentation: A Systematic Literature Review and a Series of Experiments. 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), 187–197. https://doi.org/10.1109/MSR.2017.42

Albert, Reka, & Barabasi, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 74(1), 47–97. https://doi.org/10.1103/RevModPhys.74.47

Albert, Réka, Jeong, H., & Barabási, A.-L. (1999). Diameter of the World-Wide Web. Nature, 401(6749), 130–131. https://doi.org/10.1038/43601

Alhajj, R., & Rokne, J. (2014). Encyclopedia of Social Network Analysis and Mining. Springer Publishing Company, Incorporated.

Alonso-Magdalena, P., Quesada, I., & Nadal, A. (2011). Endocrine disruptors in the etiology of type 2 diabetes mellitus. Nature Reviews Endocrinology, 7(6), 346–353. https://doi.org/10.1038/nrendo.2011.56

Alstott, J., Bullmore, E., & Plenz, D. (2014). powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions. PLoS ONE, 9(1), e85777. https://doi.org/10.1371/journal.pone.0085777

Analyzing and modeling real-world phenomena with complex networks: A survey of applications: Advances in Physics: Vol 60, No 3. (n.d.). Retrieved March 17, 2020, from https://www.tandfonline.com/doi/full/10.1080/00018732.2011.572452?casa_token=VUtsXTeZCOAAAAAA%3AxJdfMuwzZirsr8-wwXbJHLZTzvTgfipwbcGBYxPJIzeiTD_i6hplf0HmhzLCXBSh7NsHpNOmAhUz

Association, A. D. (2000). Type 2 Diabetes in Children and Adolescents. Pediatrics, 105(3), 671–680. https://doi.org/10.1542/peds.105.3.671

Association, A. D. (2004). Gestational Diabetes Mellitus. Diabetes Care, 27(suppl 1), s88–s90. https://doi.org/10.2337/diacare.27.2007.S88

Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. Lancet, 383(9911), 69–82. https://doi.org/10.1016/S0140-6736(13)60591-7

Auber, D. (2004). Tulip—A Huge Graph Visualization Framework. In M. Jünger & P. Mutzel (Eds.), Graph Drawing Software (pp. 105–126). Springer. https://doi.org/10.1007/978-3-642-18638-7_5

Authoritative sources in a hyperlinked environment | Journal of the ACM. (n.d.). Retrieved March 17, 2020, from https://dl.acm.org/doi/abs/10.1145/324133.324140?casa_token=sTrbA8mQ4_MAAAAA:8km1VtZWLA7jk9TaD3yXPtPmZbzm3Aipb4FP_T9T2cZrX3VBbCC0PqGj5fCJzDMhoiD7GLt3KS75

Baeza-yates, R., & Castillo, C. (2002). Balancing Volume, Quality and Freshness in Web Crawling. In Soft Computing Systems - Design, Management and Applications, 565–572.

Bae, S., & Choi, J. H. (2000, April). Cyberlinks between human rights NGOs: A network analysis. Paper presented to the 58th annual national meeting of the Midwest Political Science Association, Chicago.

Barabási, A.-L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: The topology of the world-wide web. Physica A: Statistical Mechanics and Its Applications, 281(1), 69–77. https://doi.org/10.1016/S0378-4371(00)00018-2

BARABÁSI, A.-L., & BONABEAU, E. (2003). Scale-Free Networks. Scientific American, 288(5), 60–69. JSTOR.

Bastian, M., Heymann, S., & Jacomy, M. (n.d.). Gephi: An Open Source Software for Exploring and Manipulating Networks. 2.

Batagelj, V., & Mrvar, A. (n.d.). Pajek—Analysis and Visualization of Large Networks. 27.

Bergenstal, R. M., Tamborlane, W. V., Ahmann, A., Buse, J. B., Dailey, G., Davis, S. N., … Group, for the S. 3 S. (2011). Sensor-Augmented Pump Therapy for A1C Reduction

(STAR 3) Study: Results from the 6-month continuation phase. Diabetes Care, 34(11), 2403–2405. https://doi.org/10.2337/dc11-1248

Blashfield, R. K., & Aldenderfer, M. S. (1988). The Methods and Problems of Cluster Analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), Handbook of Multivariate Experimental Psychology (pp. 447–473). Springer US. https://doi.org/10.1007/978-1-4613-0893-5_14

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 2008(10), P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Blumer, I., Edelman, S. V., & Hirsch, I. B. (2012). Insulin-pump therapy for type 1 diabetes mellitus. The New England Journal of Medicine, 367(4), 383; author reply 383-384. https://doi.org/10.1056/NEJMc1206221

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, D.-U. (2006). Complex networks: Structure and dynamics. Physics Reports, 424(4), 175–308. https://doi.org/10.1016/j.physrep.2005.10.009

Boguna, M., Pastor-Satorras, R., Diaz-Guilera, A., & Arenas, A. (2003). Emergence of clustering, correlations, and communities in a social network model. ArXiv:Cond-Mat/0309263. http://arxiv.org/abs/cond-mat/0309263

Braun, V., Clarke, V., & Terry, G. (2014). Thematic Analysis (pp. 95–113). https://doi.org/10.1007/978-1-137-29105-9_7

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3 Data-Driven Documents. IEEE Transactions on Visualization and Computer Graphics, 17(12), 2301–2309. https://doi.org/10.1109/TVCG.2011.185

Boukacem-Zeghmouri, C., & Schöpfel, J. (2013). 9 - Beyond the Google generation: Towards community-specific usage patterns of scientific information. In D. Baker & W. Evans (Eds.), Trends, Discovery, and People in the Digital Age (pp. 137–151). Chandos Publishing. https://doi.org/10.1016/B978-1-84334-723-1.50009-7

Brandes, U. (2001). A faster algorithm for betweenness centrality. The Journal of Mathematical Sociology, 25(2), 163–177. https://doi.org/10.1080/0022250X.2001.9990249

Brankston, G. N., Mitchell, B. F., Ryan, E. A., & Okun, N. B. (2004). Resistance exercise decreases the need for insulin in overweight women with gestational diabetes mellitus. American Journal of Obstetrics and Gynecology, 190(1), 188–193. https://doi.org/10.1016/S0002-9378(03)00951-7

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Burt, R. S. (1976). Positions in Networks. Social Forces, 55(1), 93–122. https://doi.org/10.1093/sf/55.1.93

Card, M. (1999). Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann.

Cardoso, S., & Charlet, J. (n.d.). L'anonymisation au service du repérage conceptuel dans le contexte de la SLA. 15.

Castillo, C. (2005). Effective web crawling. ACM SIGIR Forum, 39(1), 55. https://doi.org/10.1145/1067268.1067287

Castellví, M. T. C. (1999). Terminology: Theory, methods and applications. John Benjamins Publishing.

Ceci, M., & Malerba, D. (2007). Classifying web documents in a hierarchy of categories: A comprehensive study. Journal of Intelligent Information Systems, 28(1), 37–78. https://doi.org/10.1007/s10844-006-0003-2

Chan, J. C. N., Malik, V., Jia, W., Kadowaki, T., Yajnik, C. S., Yoon, K.-H., & Hu, F. B. (2009). Diabetes in Asia: Epidemiology, Risk Factors, and Pathophysiology. JAMA, 301(20), 2129–2140. https://doi.org/10.1001/jama.2009.726

Chen, C. (2010). Information visualization. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 387–403. https://doi.org/10.1002/wics.89

Chen, M., Kuzmin, K., & Szymanski, B. K. (2014). Community Detection via Maximization of Modularity and Its Variants. IEEE Transactions on Computational Social Systems, 1(1), 46–65. https://doi.org/10.1109/TCSS.2014.2307458

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice, 138, 271–281. https://doi.org/10.1016/j.diabres.2018.02.023

Cichocki, A., & Phan, A.-H. (2009). Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations. IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences, E92-A(3), 708–721.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. Physical Review E, 70(6), 066111. https://doi.org/10.1103/PhysRevE.70.066111

Collective dynamics of 'small-world' networks | Nature. (n.d.). Retrieved March 17, 2020, from https://www.nature.com/articles/30918.

Complex brain networks: Graph theoretical analysis of structural and functional systems | Nature Reviews Neuroscience. (n.d.). Retrieved March 17, 2020, from https://www.nature.com/articles/nrn2575?message=remove&lang=en

Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in web search. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 407–416. https://doi.org/10.1145/1240624.1240690

Deibert, R., Palfrey, J., Rohozinski, R., & Zittrain, J. (2010). Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace. MIT Press.

DiBiase, D., MacEachren, A. M., Krygier, J. B., & Reeves, C. (1992). Animation and the Role of Map Design in Scientific Visualization. Cartography and Geographic Information Systems, 19(4), 201–214. https://doi.org/10.1559/152304092783721295

Dramé, K., Diallo, G., Delva, F., Dartigues, J. F., Mouillet, E., Salamon, R., & Mougin, F. (2014). Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: An application to Alzheimer's disease. Journal of Biomedical Informatics, 48, 171–182. https://doi.org/10.1016/j.jbi.2013.12.013

Easley, D., & Kleinberg, J. (n.d.). Networks, Crowds, and Markets: 833.

ECT volume 26 issue 5 Cover and Back matter. (2010). Econometric Theory, 26(5), b1–b4. https://doi.org/10.1017/S0266466609990685

El-Sappagh, S., & Ali, F. (2016). DDO: A diabetes mellitus diagnosis ontology. Applied Informatics, 3(1), 5. https://doi.org/10.1186/s40535-016-0021-2

Enquist, B. J., Brown, J. H., & West, G. B. (n.d.). Allometric scaling of plant energetics and population density. 7.

Eysenbach, G., & Kohler, Ch. (2003). What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet. AMIA Annual Symposium Proceedings, 2003, 225–229.

Fiedler, M. (n.d.). ALGEBRAIC CONNECTIVITY OF GRAPHS*). 9.

Fox, S. (n.d.). The Engaged E-patient Population. 4.

Fox, S., & Fallows, D. (2003). Internet Health Resources (SSRN Scholarly Paper No. ID 2054071). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=2054071

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., & Treinen, M. (2001). What makes Web sites credible? A report on a

large quantitative study. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 61–68. https://doi.org/10.1145/365024.365037

Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. Quality and Quantity, 14(4), 585–592. https://doi.org/10.1007/BF00184720

Fursin, G., Memon, A., Guillon, C., & Lokhmotov, A. (2015). Collective Mind, Part II: Towards Performance- and Cost-Aware Software Engineering as a Natural Science. ArXiv:1506.06256 [Cs]. http://arxiv.org/abs/1506.06256

Gaharwar, R. D., & Shah, D. B. (2018). Blackhat Search Engine Optimization Techniques (SEO) and Counter Measures. https://doi.org/10.32628/ijsrst1840117

Ghemawat, P. (n.d.). THE SNOWBALL EFFECT. 17.

Granka, L., Feusner, M., & Lorigo, L. (n.d.). Eyetracking in Online Search. 28.

Guariguata, L., Whiting, D. R., Hambleton, I., Beagley, J., Linnenkamp, U., & Shaw, J. E. (2014). Global estimates of diabetes prevalence for 2013 and projections for 2035. Diabetes Research and Clinical Practice, 103(2), 137–149. https://doi.org/10.1016/j.diabres.2013.11.002

Guest, G., MacQueen, K., & Namey, E. (2012). Applied Thematic Analysis. SAGE Publications, Inc. https://doi.org/10.4135/9781483384436

Hagberg, A., Swart, P., & S Chult, D. (2008). Exploring network structure, dynamics, and function using networkx (LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab. (LANL), Los Alamos, NM (United States). https://www.osti.gov/biblio/960616

Hallgren, K. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. Tutorials in Quantitative Methods for Psychology, 8, 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Haw, J. S., Galaviz, K. I., Straus, A. N., Kowalski, A. J., Magee, M. J., Weber, M. B., Ali, M. K. (2017). Long-term Sustainability of Diabetes Prevention Approaches: A Systematic Review and Meta-analysis of Randomized Clinical Trials. JAMA Internal Medicine, 177(12), 1808–1817. https://doi.org/10.1001/jamainternmed.2017.6040

Hirsch, I. B. (2009). Realistic Expectations and Practical Use of Continuous Glucose Monitoring for the Endocrinologist. The Journal of Clinical Endocrinology & Metabolism, 94(7), 2232–2238. https://doi.org/10.1210/jc.2008-2625

Hirsch, I. B. (2012). Low Glucose Suspend: Ready for Prime Time? Diabetes Technology & Therapeutics, 14(3), 201–202. https://doi.org/10.1089/dia.2012.0036

Höchstötter, N., & Lewandowski, D. (2009). What users see – Structures in search engine results pages. Information Sciences, 179(12), 1796–1812. https://doi.org/10.1016/j.ins.2009.01.028

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online Learning for Latent Dirichlet Allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), Advances in Neural Information Processing Systems 23 (pp. 856–864). Curran Associates, Inc. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf

Holland, P. W., & Leinhardt, S. (1971). Transitivity in Structural Models of Small Groups. Comparative Group Studies, 2(2), 107–124. https://doi.org/10.1177/104649647100200201

Hu, F. B., van Dam, R. M., & Liu, S. (2001). Diet and risk of Type II diabetes: The role of types of fat and carbohydrate. Diabetologia, 44(7), 805–817. https://doi.org/10.1007/s001250100547

Integration and search engine bias—Cornière—2014—The RAND Journal of Economics—Wiley Online Library. (n.d.). Retrieved February 12, 2020, from https://onlinelibrary.wiley.com/doi/full/10.1111/1756-2171.12063

International Diabetes Federation. (2005). Global guideline for type 2 diabetes. Retrieved from http://library.imf.org/Restricted/docs/IDF_GlobalGuidelineForType2Diabetes.pdf

Ibrahim, M. (2012). THEMATIC ANALYSIS: A CRITICAL REVIEW OF ITS PROCESS AND EVALUATION. 1(1), 9.

Jackson, M. H. (1997). Assessing the Structure of Communication on the World Wide Web. Journal of Computer-Mediated Communication, 3(1). https://doi.org/10.1111/j.1083-6101.1997.tb00063.x

Jacomy, M., Girard, P., Ooghe-Tabanou, B., & Venturini, T. (n.d.). Hyphe, a Curation-Oriented Approach to Web Crawling for the Social Sciences. 4.

Jacomy, M., Venturini, T., Heymann, S., & Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLOS ONE, 9(6), e98679. https://doi.org/10.1371/journal.pone.0098679

Julien, G., Tayeb, M., F, S. L., Catherine, L., Jean, C., N, R. P., & J, D. S. (2013). Integrating the Human Phenotype Ontology into HeTOP Terminology-Ontology Server. Studies in Health Technology and Informatics, 961–961. https://doi.org/10.3233/978-1-61499-289-9-961

Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. Physical Review E, 83(1), 016107. https://doi.org/10.1103/PhysRevE.83.016107

Keim, D. A. (2002). Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1), 1–8. https://doi.org/10.1109/2945.981847

Kent, D., D'Eramo Melkus, G., Stuart, P. "Mickey" W., McKoy, J. M., Urbanski, P., Boren, S. A., … Lipman, R. (2013). Reducing the Risks of Diabetes Complications Through Diabetes Self-Management Education and Support. Population Health Management, 16(2), 74–81. https://doi.org/10.1089/pop.2012.0020

Kernighan, B. W., & Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. The Bell System Technical Journal, 49(2), 291–307. https://doi.org/10.1002/j.1538-7305.1970.tb01770.x

Kim, J., Thomas, P., Sankaranarayana, R., Gedeon, T., & Yoon, H.-J. (2015). Eye-tracking analysis of user behavior and performance in web search on large and small screens. Journal of the Association for Information Science and Technology, 66(3), 526–544. https://doi.org/10.1002/asi.23187

Kleinberg, J., & Lawrence, S. (2001). The Structure of the Web. Science, 294(5548), 1849–1850. https://doi.org/10.1126/science.1067014

KLEINBERG, J. M. (n.d.). Authoritative Sources in a Hyperlinked Environment. 29.

Kliman-Silver, C., Hannak, A., Lazer, D., Wilson, C., & Mislove, A. (2015). Location, Location, Location: The Impact of Geolocation on Web Search Personalization. Proceedings of the 2015 Internet Measurement Conference, 121–127. https://doi.org/10.1145/2815675.2815714

Knowler, W. C., Barrett-Connor, E., Fowler, S. E., Hamman, R. F., Lachin, J. M., Walker, E. A., & Nathan, D. M. (2002). Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England Journal of Medicine, 346(6), 393–403. https://doi.org/10.1056/NEJMoa012512

Kobourov, S. G. (2012). Spring Embedders and Force Directed Graph Drawing Algorithms. ArXiv:1201.3011 [Cs]. http://arxiv.org/abs/1201.3011

Krebs, V. (1999). Working in the Connected World: 5.

Kuske, S., Schiereck, T., Grobosch, S., Paduch, A., Droste, S., Halbach, S., & Icks, A. (2017). Diabetes-related information-seeking behaviour: A systematic review. Systematic Reviews, 6(1). https://doi.org/10.1186/s13643-017-0602-8

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. Physical Review E, 84(6), 066122. https://doi.org/10.1103/PhysRevE.84.066122

Lawson, T. (2004). A conception of ontology.

Learning About Information Technologies and Social Change: The Contribution of Social Informatics: The Information Society: Vol 16, No 3. (n.d.). Retrieved March 17, 2020, from https://www.tandfonline.com/doi/abs/10.1080/01972240050133661

Major, C. A., Henry, M. J., de Veciana, M., & Morgan, M. A. (1998). The Effects of Carbohydrate Restriction in Patients With Diet-Controlled Gestational Diabetes. Obstetrics & Gynecology, 91(4), 600–604. https://doi.org/10.1016/S0029-7844(98)00003-9

Manning, C. D., Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Martin, S., Schneider, B., Heinemann, L., Lodwig, V., Kurth, H.-J., Kolb, H., … for the ROSSO Study Group. (2006). Self-monitoring of blood glucose in type 2 diabetes and long-term outcome: An epidemiological cohort study. Diabetologia, 49(2), 271–278. https://doi.org/10.1007/s00125-005-0083-5

Mayer-Davis, E. J., Rifas-Shiman, S. L., Zhou, L., Hu, F. B., Colditz, G. A., & Gillman, M. W. (2006). Breast-Feeding and Risk for Childhood Obesity: Does maternal diabetes or obesity status matter? Diabetes Care, 29(10), 2231–2237. https://doi.org/10.2337/dc06-0974

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics, 10(1), 213. https://doi.org/10.1186/1471-2105-10-213

METZGER, B. eds. (1998). Proceedings of the Fourth International Workshop Conference on Gestational Diabetes Mellitus. Diabetes Care, 21(2), B1–B167.

Morahan-Martin, J. M. (2004). How Internet Users Find, Evaluate, and Use Online Health Information: A Cross-Cultural Review. CyberPsychology & Behavior, 7(5), 497–510. https://doi.org/10.1089/cpb.2004.7.497

Mjølstad, B. (2015). Knowing patients as persons. A theory-driven, qualitative study of the relevance of person-related knowledge in primary health care.

Models and Methods in Social Network Analysis. (n.d.). Retrieved April 5, 2019, from https://www.goodreads.com/work/best_book/377645-models-and-methods-in-social-network-analysis-structural-analysis-in-th

Monmonier, M. (1985). Review of Semiology of Graphics: Diagrams, Networks, Maps., , ; The Visual Display of Quantitative Information [Review of Review of Semiology of Graphics: Diagrams, Networks, Maps., , ; The Visual Display of Quantitative Information, by J. Bertin, W. J. Berg, & E. R. Tufte]. Annals of the Association of American Geographers, 75(4), 605–609. JSTOR.

Murray, D. G. (2013). Tableau Your Data! Fast and Easy Visual Analysis with Tableau Software. John Wiley & Sons.

Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research—Fisher—2016—Intelligent Systems in Accounting, Finance and Management—Wiley Online Library. (n.d.). Retrieved March 17,2020,from https://onlinelibrary.wiley.com/doi/full/10.1002/isaf.1386?casa_token=nNg8j4ZXZ78AAAAA%3Ajl7M37FMBOxYBVe6HDpftk6kZDWJMsVrLMOErCdQUC-w-mV5d_67KDU6HbV-dOdLs9KNHtbPBDksKJA

Newman, M., Barabási, A.-L., & Watts, D. J. (2011). The Structure and Dynamics of Networks. Princeton University Press.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98(2), 404–409. https://doi.org/10.1073/pnas.98.2.404

Newman, M. E. J. (2004). Detecting community structure in networks. The European Physical Journal B, 38(2), 321–330. https://doi.org/10.1140/epjb/e2004-00124-y

Newman, M. E. J. (2006). Modularity and community structure in networks. Proceedings of the National Academy of Sciences, 103(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. Physical Review E, 69(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nottingham <mnot@mnot.net>, M. (n.d.). Well-Known Uniform Resource Identifiers (URIs). Retrieved March 17, 2020, from https://tools.ietf.org/html/rfc8615

NW, 1615 L. St, Suite 800Washington, & Inquiries, D. 20036USA202-419-4300 | M.-857-8562 | F.-419-4372 | M. (2012, March 9). Search Engine Use 2012. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2012/03/09/search-engine-use-2012-2/

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November 11). The PageRank Citation Ranking: Bringing Order to the Web. [Techreport]. Stanford InfoLab. http://ilpubs.stanford.edu:8090/422/

Palmer, J. W., Bailey, J. P., & Faraj, S. (2000). The Role of Intermediaries in the Development of Trust on the Www: The Use and Prominence of Trusted Third Parties and Privacy Statements. Journal of Computer-Mediated Communication, 5(3). https://doi.org/10.1111/j.1083-6101.2000.tb00342.x

Park, H. W. (n.d.). Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web. 14.

Park, H. W. (2002, November 4). Examining the determinants of who is hyperlinked to whom (1996 - 2002) [Text]. First Monday, ISSN 1396-0466. https://firstmonday.org/ojs/index.php/fm/article/download/1005/926?inline=1

Park, H. W., & Thelwall, M. (2003). Hyperlink Analyses of the World Wide Web: A Review. Journal of Computer-Mediated Communication, 8(4). https://doi.org/10.1111/j.1083-6101.2003.tb00223.x

Pavlopoulos, G. A., Paez-Espino, D., Kyrpides, N. C., & Iliopoulos, I. (2017). Empirical Comparison of Visualization Tools for Larger-Scale Network Analysis [Review Article]. Advances in Bioinformatics. https://doi.org/10.1155/2017/1278932

Peixoto, T. P. (2014). Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. Physical Review E, 89(1), 012804. https://doi.org/10.1103/PhysRevE.89.012804

Pfaender, F., Jacomy, M., & Fouetillou, G. (2006). Two Visions of the Web: From Globality to Localities. 2006 2nd International Conference on Information Communication Technologies, 1, 566–571. https://doi.org/10.1109/ICTTA.2006.1684433

Popejoy, L. L., Khalilia, M. A., Popescu, M., Galambos, C., Lyons, V., Rantz, M., Hicks, L., & Stetzer, F. (2015). Quantifying care coordination using natural language processing and domain-specific ontology. Journal of the American Medical Informatics Association, 22(e1), e93–e103. https://doi.org/10.1136/amiajnl-2014-002702

Postel, J. (n.d.). Internet Protocol. Retrieved March 17, 2020, from https://tools.ietf.org/html/rfc791

Pothen, A., Simon, H. D., & Liou, K.-P. (1990). Partitioning Sparse Matrices with Eigenvectors of Graphs. SIAM Journal on Matrix Analysis and Applications, 11(3), 430–452. https://doi.org/10.1137/0611030

Purcell, K., Brenner, J., & Rainie, L. (n.d.). Even though online Americans are more satisfied than ever with the performance of search engines, strong majorities have negative views of personalized search results and targeted ads. 42.

Purchase, H. C., Cohen, R. F., & James, M. (1996). Validating graph drawing aesthetics. In F. J. Brandenburg (Ed.), Graph Drawing (pp. 435–446). Springer. https://doi.org/10.1007/BFb0021827

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., … Makaroff, L. E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Research and Clinical Practice, 128, 40–50. https://doi.org/10.1016/j.diabres.2017.03.024

Olokoba, A. B., Obateru, O. A., & Olokoba, L. B. (2012). Type 2 Diabetes Mellitus: A Review of Current Trends. Oman Medical Journal, 27(4), 269–273. https://doi.org/10.5001/omj.2012.68

Organization, W. H. (2016). World Health Statistics 2016: Monitoring Health for the SDGs Sustainable Development Goals. World Health Organization.

Oser, T. K., Oser, S. M., McGinley, E. L., & Stuckey, H. L. (2017). A Novel Approach to Identifying Barriers and Facilitators in Raising a Child With Type 1 Diabetes: Qualitative Analysis of Caregiver Blogs. JMIR Diabetes, 2(2), e27. https://doi.org/10.2196/diabetes.8966

Poolsup, N., Suksomboon, N., & Rattanasookchit, S. (2009). Meta-Analysis of the Benefits of Self-Monitoring of Blood Glucose on Glycemic Control in Type 2 Diabetes Patients: An Update. Diabetes Technology & Therapeutics, 11(12), 775–784. https://doi.org/10.1089/dia.2009.0091

Ramachandran, A., Mary, S., Yamuna, A., Murugesan, N., & Snehalatha, C. (2008). High Prevalence of Diabetes and Cardiovascular Risk Factors Associated With Urbanization in India. Diabetes Care, 31(5), 893–898. https://doi.org/10.2337/dc07-1207

Ramos, J. (n.d.). Using TF-IDF to Determine Word Relevance in Document Queries. 4.

Reas, C., & Fry, B. (2004). Processing.org: Programming for artists and designers. ACM SIGGRAPH 2004 Web Graphics, 3. https://doi.org/10.1145/1186194.1186198

Roglic, G., & World Health Organization (Eds.). (2016). Global report on diabetes. Geneva, Switzerland: World Health Organization.

Sabidussi, G. (1966). The centrality index of a graph. Psychometrika, 31(4), 581–603. https://doi.org/10.1007/BF02289527

Schaefer-Graf, U. M., Hartmann, R., Pawliczak, J., Passow, D., Abou-Dakn, M., Vetter, K., & Kordonouri, O. (2006). Association of Breast-feeding and Early Childhood Overweight in Children From Mothers With Gestational Diabetes Mellitus. Diabetes Care, 29(5), 1105–1107. https://doi.org/10.2337/dc05-2413

Scott, J. (1988). Social Network Analysis. Sociology, 22(1), 109–127. https://doi.org/10.1177/0038038588022001007

Semantic web or web of data? A diachronic study (1999 to 2017) of the publications of tim berners-lee and the world wide web consortium—Machado—2019—Journal of the Association for Information Science and Technology—Wiley Online Library. (n.d.). Retrieved March 17, 2020, from https://asistdl.onlinelibrary.wiley.com/doi/full/10.1002/asi.24111?casa_token=OtHv8088KMkAAAAA%3AnJIzGVMvsNPOqTp5Icx4s5uLhQD07KnGG-AnNwrLPLvta8EEJDEM2QgSE3zQchbD5ev-P3peRv9vrRU

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The Semantic Web Revisited. IEEE Intelligent Systems, 21(3), 96–101. https://doi.org/10.1109/MIS.2006.62

Shaw, J. E., Sicree, R. A., & Zimmet, P. Z. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Research and Clinical Practice, 87(1), 4–14. https://doi.org/10.1016/j.diabres.2009.10.007

Shen, X., Tan, B., & Zhai, C. (2007). Privacy protection in personalized search. ACM SIGIR Forum, 41(1), 4–17. https://doi.org/10.1145/1273221.1273222

Shiffrin, R. M., & Börner, K. (2004). Mapping knowledge domains. Proceedings of the National Academy of Sciences, 101(suppl 1), 5183–5185. https://doi.org/10.1073/pnas.0307852100

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. Proceedings 1996 IEEE Symposium on Visual Languages, 336–343. https://doi.org/10.1109/VL.1996.545307

Sibal, L., & Home, P. D. (2009). Management of type 2 diabetes: NICE guidelines. Clinical Medicine, 9(4), 353–357. https://doi.org/10.7861/clinmedicine.9-4-353

Skovlund, S. E., & Peyrot, M. (2005). The Diabetes Attitudes, Wishes, and Needs (DAWN) Program: A New Approach to Improving Outcomes of Diabetes Care. Diabetes Spectrum, 18(3), 136–142. https://doi.org/10.2337/diaspect.18.3.136

Skupin, A. (2002). On Geometry and Transformation in Map-Like Information Visualization. In K. Börner & C. Chen (Eds.), Visual Interfaces to Digital Libraries (pp. 161–170). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-36222-3_12

Spence, I., Wainer, H., & Wainer, H. (2017, January 12). William Playfair and the invention of statistical graphs. Information Design. https://www.taylorfrancis.com/

Sporns, O., Chialvo, D. R., Kaiser, M., & Hilgetag, C. C. (2004). Organization, development and function of complex brain networks. Trends in Cognitive Sciences, 8(9), 418–425. https://doi.org/10.1016/j.tics.2004.07.008

Srinivas, A., & Velusamy, R. L. (2015). Identification of influential nodes from social networks based on Enhanced Degree Centrality Measure. 2015 IEEE International Advance Computing Conference (IACC), 1179–1184. https://doi.org/10.1109/IADCC.2015.7154889

Stacey, P. A. (2019). ECRM 2019 18th European Conference on Research Methods in Business and Management. Academic Conferences and publishing limited.

Statistical Report on Internet Development in China. The 44th Statistical Report on Internet Development in China, China Internet Network Information Center. 2019; https://cnnic.com.cn/IDR/ReportDownloads/201911/P020191112539794960687.pdf.

Susan van, D., Beulens, J. W. J., Yvonne T. van der, S., Grobbee, D. E., & Nealb, B. (2010). The global burden of diabetes and its complications: An emerging pandemic. European Journal of Cardiovascular Prevention & Rehabilitation, 17(1_suppl), s3–s8. https://doi.org/10.1097/01.hjr.0000368191.86614.5a

Telea, A. C. (2014). Data Visualization: Principles and Practice, Second Edition. CRC Press.

The AfD's Facebook Wall as a Hub for Right-Wing Mobilisation in Germany. (2015, August 28). Kai Arzheimer. http://www.kai-arzheimer.com/my-apsa-2015-paper-the-afds-facebook-wall-as-a-hub-for-right-wing-mobilisation-in-germany/

The expert patients programme online, a 1-year study of an Internet-based self-management programme for people with long-term conditions—Kate R. Lorig, Philip L. Ritter, Ayesha Dost, Kathryn Plant, Diana D. Laurent, Ian Mcneil, 2008. (n.d.). Retrieved September 25, 2019, from https://journals.sagepub.com/doi/abs/10.1177/1742395308098886

Todd, J. A. (2010). Etiology of Type 1 Diabetes. Immunity, 32(4), 457–467. https://doi.org/10.1016/j.immuni.2010.04.001

Thomas, D. R. (2003). A general inductive approach for qualitative data analysis. 11.

Tufte, E. R. (1983). The visual display of quantitative information. https://doi.org/10.1097/01445442-198507000-00012

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. Psychological Bulletin, 101(1), 140–146. https://doi.org/10.1037/0033-2909.101.1.140

Uys, J. W., du Preez, N. D., & Uys, E. W. (2008). Leveraging unstructured information using topic modelling. PICMET '08 - 2008 Portland International Conference on Management of Engineering Technology, 955–961. https://doi.org/10.1109/PICMET.2008.4599703

Venturini, T., Jacomy, M., Bounegru, L., & Gray, J. (2017). Visual Network Exploration for Data Journalists (SSRN Scholarly Paper ID 3043912). Social Science Research Network. https://papers.ssrn.com/abstract=3043912

Whiting, D. R., Guariguata, L., Weil, C., & Shaw, J. (2011). IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes Research and Clinical Practice, 94(3), 311–321. https://doi.org/10.1016/j.diabres.2011.10.029

Wills, R. S. (2006). Google's pagerank. The Mathematical Intelligencer, 28(4), 6–11. https://doi.org/10.1007/BF02984696

Yang, W., Lu, J., Weng, J., Jia, W., Ji, L., Xiao, J., He, J. (2010). Prevalence of Diabetes among Men and Women in China. New England Journal of Medicine, 362(12), 1090–1101. https://doi.org/10.1056/NEJMoa0908292

Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. Journal of Anthropological Research, 33(4), 452–473. https://doi.org/10.1086/jar.33.4.3629752

Zhou, H. (2003). Distance, dissimilarity index, and network community structure. Physical Review E, 67(6), 061901. https://doi.org/10.1103/PhysRevE.67.061901

List of Figures

List of Tables