Analyse d'Articles d'Ingénierie des Connaissances : Session Web Social de l'ISWC 2011

Matti Schneider-Ghibaudo

Knowledge and Information Systems
Département Sciences Informatiques
École Polytechnique Universitaire Polytech'Nice-Sophia Antipolis
Sophia-Antipolis, France

ghibaudo@polytech.unice.fr

Abstract. Ce document s'intéresse à deux articles publiés lors de l'International Semantic Web Conference (ISWC) 2011. Son objectif est de résumer les méthodes et découvertes de chacun d'entre eux, puis d'en souligner les limites et biais, avant de conclure quant aux éléments communs et contradictoires les réunissant ou les opposant. Nous verrons ainsi que l'augmentation sémantique de contenu a une efficacité démontrée pour l'amélioration d'un processus de recherche à facettes, mais pas pour la détermination d'expertise dans un groupe.

Mots-clés : social, sémantique, extraction de connaissances, recherche à facettes, Twitter, email, échanges informels

1. Introduction

Les deux articles que j'ai choisi pour cette analyse proviennent de la session “Social Web” de l'International Web Semantic Conference 2011. Ces articles s'intéressent tous deux à la manière dont la connaissance peut être extraite à partir d'échanges informels, grâce à l'augmentation sémantique du contenu. Ils reposent notamment sur la création de profils utilisateurs pour améliorer l'efficacité de la découverte d'information. Certains points importants les différencient néanmoins.

Tout d'abord, leur objectif n'est pas le même. En effet, l'un se concentre sur la recherche de contenu, tandis que l'autre s'intéresse à la création d'une topologie de l'expertise au sein d'une organisation. Les objets d'expérimentation ne sont donc bien évidemment pas les mêmes. Ensuite, les données source sont différentes : dans le premier cas, il s'agit de messages du réseau social Twitter (*tweet*), dans l'autre de courriels (*email*). Il est néanmoins intéressant de rapprocher ces articles, ne serait-ce que pour leur fort contraste méthodologique et rédactionnel. J'ai en effet été fort surpris de la disparité de clarté ressentie, et plus encore de la découverte d'incohérences et limitations méthodologiques sévères pour l'un d'entre eux.

Nous nous attacherons donc dans un premier temps à résumer la problématique soulevée et à éclaircir l'approche choisie et les solutions proposées par chacun des articles. Ensuite, nous examinerons les limitations et biais parfois rencontrés. Enfin, nous réunirons dans la mesure du possible les deux documents pour en extraire les éléments méthodologiques communs liés aux techniques sémantiques.

2. Recherche à Facettes Adaptative sur Twitter

2.1. Définitions

La **recherche à facettes** est une méthode de recherche d'information qui consiste à décomposer les résultats d'une recherche selon différents critères, ou *facettes*. Il s'agit d'une construction plus complexe que la simple recherche selon des types de métadonnées, puisque les facettes peuvent également être directement liées au contenu. Par exemple, sont des facettes valides la présence de certains mots dans le corps d'un texte, ou encore la référence à certaines entités ou concepts. La valeur à laquelle est faite la référence est simplement appelée **valeur** associée à la facette. Par exemple, pour la facette “présence d'un mot-clé”, on peut avoir autant de valeurs pour cette facette que de mots-clés. Le couple (“présence d'un hashtag”, “#iwsc2011”) constitue ainsi un couple facette-valeur dont la valeur est “#iwsc2011”.

Un **tweet** est un message textuel de moins de 140 caractères, pouvant éventuellement contenir des URLs pointant vers des ressources web. Un tweet possède un certain nombre de métadonnées, telles que son auteur ou sa date d'émission, mais ne possède pas de facettes explicites. Le langage et la syntaxe utilisés dans ces messages ne sont pas standard ; la limitation de taille pousse en effet les utilisateurs à utiliser des artifices syntaxiques tels qu'abréviations, mais également *hashtags* ou *mentions*.

Un **hashtag** est un mot-clé précédé d'un glyphe dièse (#), qui permet de regrouper des tweets dans l'équivalent d'un fil de discussion décentralisé et d'en faciliter la recherche. Une **mention** est l'appellation donnée à un élément syntaxique constitué d'une arobase (@) suivie d'un nom d'utilisateur du service Twitter. Elle permet ainsi de faire facilement référence à une entité de ce réseau social. Un **retweet** est la retransmission d'un tweet par un autre utilisateur du réseau que son auteur original. Une telle retransmission est généralement signe d'assentiment ou d'appui.

2.2. Problématique

La recherche dans Twitter est actuellement limitée, puisque les seuls filtres disponibles sont la présence de hashtags, de mentions ou de mots-clés. Les résultats ne sont organisés que par date d'émission, et n'ont aucune notion de pertinence au-delà du nombre de retweets, événement somme toute assez rare. L'objectif de cet article est donc (i) de déterminer une méthode de recherche permettant à l'utilisateur d'aboutir au résultat souhaité le plus rapidement possible pour une requête donnée, et (ii) de valider expérimentalement l'hypothèse selon laquelle l'augmentation sémantique du contenu permet d'améliorer cette pertinence. On remarquera qu'il n'est pas aisé d'extraire des données sémantiques depuis des tweets, à cause notamment des contraintes de taille et de syntaxe définies plus haut.

2.3. Solution Apportée

Les auteurs utilisent deux moyens différents pour ajouter des valeurs de facettes aux tweets : l'**enrichissement de tweet** (*tweet-based enrichment*) et l'**enrichissement de lien** (*link-based enrichment*). Dans les deux cas, il s'agit d'ajouter des informations sémantiques à un tweet, telles qu'entités et concepts mentionnés. L'enrichissement de tweet, se contente de procéder sur le contenu du tweet seul. Dans le cas de l'enrichissement de lien, on accède en plus aux éventuelles URLs citées dans le tweet, et les informations extraites des documents liés lui sont associées. À partir des informations sémantiques ainsi associées à un tweet, une recherche à facettes adaptative est mise en place. L'adjectif *adaptatif* fait référence à la création d'un profil de recherche dépendant de l'utilisateur et du contexte de la recherche. La méthodologie exacte de personnalisation des résultats fait partie de l'expérimentation, tout comme une comparaison de l'efficacité des deux méthodes d'enrichissement.

Attention néanmoins : l'article ne traite pas de l'organisation des *résultats* de la recherche eux-mêmes, c'est-à-dire de la présentation des tweets répondant à certains critères de recherche (cette question est en partie traitée par [3]). Il s'agit ici de déterminer le tri le plus pertinent pour les *facettes* (et leurs valeurs), pour permettre à l'utilisateur de raffiner le plus rapidement possible sa requête. Les tweets répondant aux critères sont eux-mêmes classés uniquement par ancienneté. Quatre métriques sont envisagées par les auteurs pour classer les facettes. Nous allons les définir et analyser les résultats d'efficacité relative obtenus. Récapitulons tout d'abord l'architecture générale et les zones sur lesquelles les auteurs expérimentent.

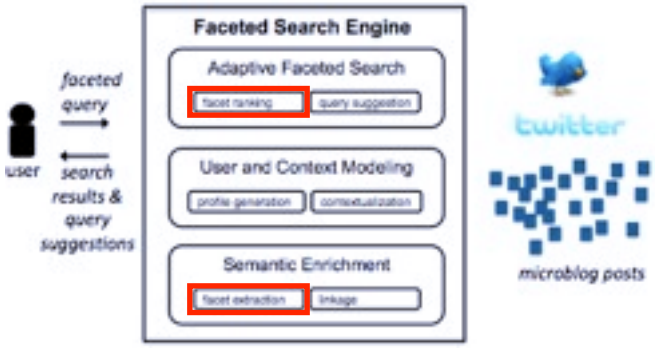

Figure 1. Architecture du moteur de recherche ; en rouge, les expérimentations qualitatives

2.4. Méthodes de Tri des Couples Facette-Valeur

Fréquence d'apparition. Cette métrique relève l'ensemble des couples facette-valeur rencontrés dans les tweets candidats, et en compte la fréquence d'apparition. Ces couples sont ensuite présentés à l'utilisateur par ordre décroissant de fréquence d'apparition. Cela permet de minimiser le risque d'ignorer une valeur importante, mais pose un problème d'efficacité. En effet, par définition, les valeurs les mieux classées sont celles qui sont le plus présentes. Elles sont donc peu discriminantes, et il faudra plus de sélections pour réduire le nombre de candidats qu'avec d'autres méthodes.

Personnalisation. Cette méthode classe les valeurs par le poids qu'elles ont dans le “profil” de l'utilisateur. Ce profil est automatiquement généré à partir de l'enrichissement sémantique des tweets publiés par l'utilisateur. Le poids de chaque couple facette-valeur est calculé par un décompte de l'occurrence de chaque couple dans l'historique de publication suivi d'une normalisation selon le plus grand nombre d'apparitions. Une valeur absente du profil se voit attribuer un poids nul.

Diversification. Cette méthode privilégie la minimisation du nombre de choix de valeurs nécessaires à l'aboutissement de la recherche plutôt que la pertinence des facettes en elles-mêmes. L'idée est en effet de proposer des couples facette-valeur apparaissant dans le plus grand nombre de tweets possibles, mais tels que les tweets ainsi potentiellement sélectionnés ne contiennent pas des couples déjà choisis pour être présentés à l'utilisateur. Ainsi, la discrimination est forte : la sélection d'un couple éliminera une grande quantité de résultats, ce qui peut en théorie réduire les étapes.

Sensibilité temporelle. Cette métrique privilégie les tweets les plus récents, en attribuant un poids plus élevé aux couples facette-valeur apparaissant dans des tweets dont la date d'émission et la plus proche de la date de recherche.

2.5. Comparaison des Méthodes

Comme on l'a vu, le travail du moteur de recherche adaptatif est de présenter des facettes de telle sorte que les couples facette-valeur que l'utilisateur inclut dans sa recherche soient les plus visibles possible (par exemple situés en premières places dans une présentation par liste). Pour cela, les auteurs définissent la métrique *Success@k*. Cette valeur est simplement la probabilité qu'un couple facette-valeur effectivement sélectionné par l'utilisateur pour raffiner sa recherche soit présent dans les *k* premiers couples affichés. Il suffit ensuite de calculer ce *S@k* pour chacune des différentes méthodes de pondération présentées plus haut pour déterminer l'effort que l'utilisateur a du fournir pour arriver au résultat attendu : plus le *S@k* est élevé, plus les facettes-valeurs pertinentes étaient visibles, moins l'utilisateur a passé de temps à chercher les couples discriminants qu'il a finalement choisis. Les auteurs ont arbitrairement choisi des valeurs de *k* de 20, 50 et 100.

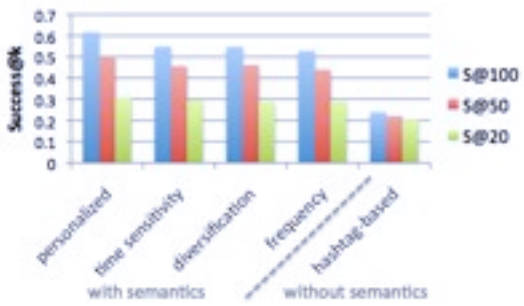

Figure 2. Comparaison des différentes méthodes d'ordonnancement des couples facette-valeur

L'efficacité relative de chaque méthode est claire dans **Fig.2** : le classement par personnalisation est de loin le meilleur, suivi d'assez près par les trois autres méthodes, la plus “basique” (basée sur la fréquence d'apparition) se classant dernière. Néanmoins, elle est toujours bien meilleure que l'utilisation d'une recherche sans enrichissement sémantique. Cette dernière observation est complétée par le calcul du MRR (*mean reciprocal rank*, le rang auquel le tweet recherché est classé).

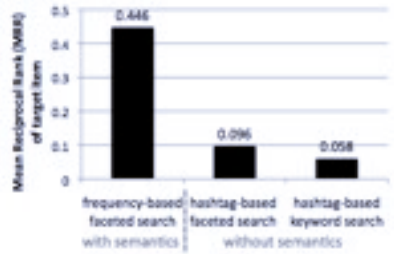

Figure 3. Comparaison des résultats de recherche avec et sans enrichissement sémantique et recherche à facettes

Ainsi, même si l'utilisation de la recherche à facettes sans augmentation sémantique prouve son utilité avec une amélioration du classement de 65%, on obtient des résultats véritablement impressionnant en utilisant la recherche à facettes sémantique : bien que la méthode employée soit la plus naïve (voir **Fig.2**), le tweet effectivement recherché est plus de 6 fois (660%) mieux placé dans la liste des candidats. Enfin, il est important de souligner que les recherches sans sémantique ne peuvent s'effectuer *que sur les tweets contenant au moins un hashtag*, alors que la recherche basée sur l'enrichissement du contenu peut s'appliquer sur n'importe quel tweet.

Terminons enfin par une comparaison des deux méthodes d'enrichissement.

Characteristics	Tweet-based enrichment	Tweet & Link-based enrichment
avg. num. of facet values per tweet	1.85	5.72
avg. num. of discoverable tweets	61161.23	75782.76
avg. num. of FVP-selects to filter results	1.95	2.25
avg. size of filtered result set	1685.320	189.48

Figure 4. Comparaison des méthodes d'enrichissement sémantique

L'utilité de l'enrichissement de liens est ici incontestable : non seulement les données sémantiques sont bien plus nombreuses (triplement du nombre de facettes par tweet), mais l'efficacité sur la recherche est impressionnante, avec une réduction d'un ordre de grandeur du nombre moyen de candidats à une requête.

2.6. Conclusion

Les auteurs ont ici bien prouvé deux points. D'une part (peut-être d'une manière particulièrement forte sur Twitter où le contenu est très court), l'efficacité de l'augmentation sémantique est très fortement améliorée par une analyse du contenu des liens. D'autre part, la recherche à facettes est très efficace sur un corpus enrichi sémantiquement, ce d'autant plus qu'une personnalisation basée sur un profil utilisateur est effectuée sur la présentation des différents couples facette-valeur.

3. Détection d'Expertise lors d'Échanges Informels

3.1. Introduction

Ce second article s'intéresse à l'extraction du *buried knowledge*, ou “savoir enfoui”, au sein d'une organisation, en utilisant les courriels échangés de manière informelle par des membres de cette organisation. Il ne cherche pas à extraire directement la connaissance des échanges, mais à repérer les experts par sujets au travers de ces conversations. L'objectif est donc de déterminer si la création de profils utilisateurs enrichis sémantiquement est efficace pour améliorer la gestion de la connaissance, en permettant de trouver des experts et d'éviter de dupliquer les compétences du groupe.

3.2. Approche

Les recherches précédentes [4] se sont attachées à déterminer les groupes d'utilisateurs partageant des intérêts communs en mesurant les fréquences relatives d'échange entre personnes. Cet article reprend une méthode appliquée sur Twitter [5], qui consiste à générer un profil utilisateur à partir de trois niveaux de sémanticité différents. Les auteurs ont ici décidé de s'intéresser uniquement aux échanges de courriel par mailing-lists, et les sources sont donc différentes (texte seul et non hashtags, par exemple). Les trois degrés de sémanticité employés sont les suivants.

Mot-clés. Le premier niveau consiste en une simple extraction de mots-clés, par la bibliothèque JATR[1]. Il n'y a donc pas ici de sémantique à proprement parler, mais un simple traitement de langue naturelle.

Entités. Le second niveau consiste à reconnaître des entités spécifiques au sein du contenu des courriels. Ces entités sont obtenues grâce au service web d'Open Calais[2].

Concepts. Le dernier niveau de sémanticité est obtenu grâce au service web Wikifier[3], qui associe des concepts définis dans Wikipédia à un texte.

---

[1] Java Automatic Term Recognition Toolkit, v1.0

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

[3] http://www.wikifyer.com/

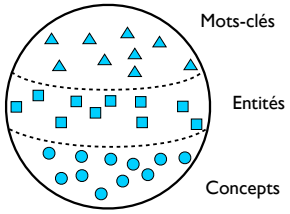

Figure 5. Modélisation d'un profil utilisateur

Ces trois degrés sont ensuite agrégés pour constituer un profil d'utilisateur contenant un ensemble de mots-clés, d'entités et de concepts représentant ses domaines d'expertise. L'idée est de permettre une exploration, ou une requête, pour isoler un expert d'un domaine spécifique dans l'organisation étudiée.

3.3. Expérimentation

La procédure expérimentale de validation des profils définie par les auteurs a été de faire remplir des questionnaires par les participants à une mailing-list, et de comparer les résultats obtenus auprès de ces humains aux résultats obtenus par une analyse d'une archive de cette mailing-list par un prototype logiciel. Le questionnaire demandait aux répondants de noter les similarités entre participants, deux à deux, sur une échelle de 1 (très différents) à 10 (très similaires). La similarité entre utilisateurs mesurée par le logiciel expérimental était calculée, elle, selon l'index de Jaccard de leurs profils. Cet index est une grandeur comprise entre 0 et 1 et indiquant la proportion d'éléments que deux ensembles partagent. Il s'agit simplement du quotient de leur intersection par leur union (voir **Fig. 6**).

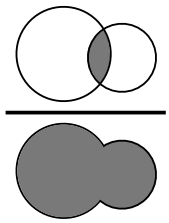

Figure 6. Symbolisation de l'index de Jaccard

La corrélation (corrélation de Pearson) entre les mesures obtenues auprès des humains et auprès du logiciel est ensuite calculée pour chacun des trois niveaux de sémanticité des profils utilisateurs, afin de confirmer ou non l'hypothèse selon laquelle l'ajout de degrés sémantiques améliore la spécificité de l'analyse. Les résultats obtenus sont consignés dans **Fig. 7**.

ID	K		NEs		Conc		Agr
	C	S	C	S	C	S	C
14	0.55	0	0.41	0.04	0.68	0	0.91
7	0.48	0.02	0.39	0.06	0.58	0	0.87
28	0.5	0.01	0.41	0.04	0.57	0	0.89
10	0.47	0.02	0.39	0.05	0.57	0	0.94
27	0.32	0.11	0.29	0.16	0.48	0.02	0.92
21	0.34	0.11	0.42	0.04	0.42	0.04	0.91
1	0.35	0.02	0.32	0.11	0.42	0.04	0.94
3	0.3	0.14	0.31	0.14	0.38	0.06	0.86
9	0.28	0.18	0.36	0.07	0.38	0.06	0.9
18	0.5	0.01	0.5	0.01	0.36	0.07	0.87
8	0.17	0.53	0.19	0.48	0.35	0.18	0.82
11	0.59	0	0.42	0.04	0.34	0.1	0.83
25	0.25	0.22	0.33	0.11	0.3	0.14	0.73
23	0.21	0.32	0.33	0.1	0.19	0.36	0.86

Figure 7. *Corrélation* (**C**) et *significance* (**S**) associée entre profils logiciels d'*utilisateurs* (colonne **ID**) et évaluation humaine, à différents niveaux de sémantique (*mots-clés* **K**, *entités* **NEs**, *concepts* **Conc**) ; l'*accord avec les autres évaluateurs* **Agr** est également donné

Ce tableau n'est pas très agréable à lire, surtout de par le manque de clarté de ses entêtes. La légende ici donnée tente de les éclairer. Précisons que l'"accord avec les autres évaluateurs" (*inter-annotator agreement*) est une mesure de l'accord d'un évaluateur individuel avec l'ensemble des autres évaluateurs. Les auteurs disent avoir encore une fois utilisé la corrélation de Pearson pour mesurer cette valeur, avec une significance inférieure à 0,001.

3.4. Conclusions

Les auteurs concluent à partir des résultats précédents que l'approche utilisée est satisfaisante, et enchaînent avec différentes visualisations des données extraites, telles que nuages de tags, ou encore graphe basé sur les forces pour isoler les zones d'expertises. Nous allons voir dans la partie suivante les critiques soulevées dans cette analyse, tant envers la méthodologie que la présentation.

4. Critique

4.1. Méthodologie

La méthodologie de validation de l'article concernant Twitter paraît saine. Elle est claire et utilise des précédents validés par les pairs, ses processus et métriques sont proprement définis. Des définitions formelles de toutes les méthodes sont offertes.

Il est en revanche difficile de faire la même observation pour le second article. Celui-ci pose en effet plusieurs problèmes méthodologiques, qui sont relativement

sérieux dans la mesure où ils empêchent la reproductibilité de l'expérience, frein majeur à l'acceptation scientifique des résultats.

Tout d'abord, le tableau de résultats (**Fig. 7**) prétend donner la corrélation entre une mesure de similarité de profils effectuée de manière logicielle par l'index de Jaccard, et une mesure de similarité par testeurs humains évaluant sur une échelle entière de 1 à 10, et où l'absence de réponse est acceptée. Néanmoins, les auteurs ne spécifient pas la manière dont les valeurs sont comparées. Formellement, nous avons donc (avec ε l'absence de réponse et ϱ la corrélation de Pearson) :

$$sim_{log} \in [0, 1]$$
$$sim_{hum} \in [\![1, 10]\!] \cup \{ \varepsilon \}$$
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \Rightarrow \varrho_{X,Y} \in \mathscr{F}([0, 1]^2, [0, 1])$$

D'où incompatibilité de domaines, mais les auteurs nous laissent sans explication quant à l'injection utilisée pour passer de Dom(sim_{hum}) à [0,1], qui est pourtant nécessaire pour pouvoir considérer sim_{hum} comme une variable aléatoire réelle (VAR) sur laquelle le coefficient de corrélation de Pearson avec sim_{log} aurait un sens.

Ensuite, les auteurs disent avoir validé l'accord entre évaluateurs (*inter-annotator agreement*) "*with Pearson correlation at significance < 0.001*". Cependant, ce calcul ne fait pas sens en tant que tel. Comme rappelé ci-dessus, la corrélation de Pearson évalue l'accord entre **deux** VAR, et non entre une et un ensemble d'autres. Après de nombreuses recherches, j'ai fini par trouver une pratique dans la littérature [6] qui consiste à déterminer la corrélation entre les notes données par un évaluateur et la moyenne des notes des autres évaluateurs. On notera que rien n'assure que cette technique soit celle employée par les auteurs, puisque rien n'est spécifié. Qui plus est, même si c'était le cas, on serait en droit de se demander pourquoi une méthode considérée comme peu adaptée à ce genre de mesures [7] est utilisée, et non une autre comme le coefficient de Fleiss, développé spécifiquement pour de tels cas [8]. L'excellente significance apparaît dès lors comme trompeuse plutôt que rassurante.

Ce genre d'imprécisions rendant la reproductibilité difficile, voire impossible, s'étend sur d'autres points tels que la mention des outils utilisés. L'extraction de concepts a ainsi été possible, d'après les auteurs, par "*the Wikify web service*", sans explication autre que la mention d'une référence. Or, cette référence [9] ne fait pas mention d'un tel service web. Elle explique effectivement les concepts d'un tel service, mais ne donne aucune information quant à un service spécifique. Il faut faire de nombreuses recherches pour réussir à aboutir à un service ressemblant à ce qui est

annoncé, Wikifier[4]. Notons néanmoins ici que le premier article commet également une erreur en ce qui concerne un service. Bien que citant correctement l'URL du service (ce qui rend possible cette critique), il est surprenant de voir que l'API citée[5] aboutit systématiquement sur un accès refusé. Une demande de détails à Twitter aboutit à la réponse suivante : "*Annotations is still more concept then reality. Maybe some day we'll have more to say about them*."[6]. Cette anomalie n'est néanmoins pas très problématique dans la mesure où l'article ne prétend pas utiliser ce service, et le cite par souci de complétude des moyens d'obtenir des métadonnées sur les tweets.

1. Données Expérimentales

L'article explorant Twitter demande à un utilisateur du réseau de retrouver un des tweets qu'il a retweeté parmi 1000 autres retweets publiés dans les dernières 24 heures. Le nombre moyen de couples facette-valeur proposés varie de 61 161 à 75 783 selon la méthode d'enrichissement choisie. Les deux métriques utilisées (classement relatif du tweet visé dans la liste des résultats par MRR et *Success@k* pour la pertinence des facettes) sont claires et les résultats sont sans ambiguïté. La base de données des tweets utilisée a été constituée par l'enregistrement de plus de 20 000 comptes, aboutissant à un corpus de plus de 30 millions de tweets. La seule mesure manquante est le nombre de testeurs ayant réellement participé à cette recherche. Une formulation ambiguë pourrait laisser entendre que tous les comptes enregistrés ont participé, mais il paraît fort peu probable qu'un tel effort de recherche ait pu être déployé sur une si grande population. Ce manque pourrait potentiellement remettre en question les résultats, mais il paraîtrait néanmoins surprenant que les auteurs, après avoir aussi soigneusement précisé leurs méthodes et étendues, n'aient pas pris la précaution de prendre une quantité convenable de testeurs.

Inversement, l'article concernant l'analyse de courriels fournit bien le nombre de testeurs, mais ne convainc pas du bien-fondé des résultats obtenus. Le corpus est constitué de 1 001 emails recueillis en 10 mois sur l'archive de la mailing-list du département *Information and Knowledge* de l'université de Sheffield[7]. Bien que ce nombre paraisse suffisamment important pour valider la démarche, on réalise un peu plus tard que tous ces échanges ne réunissent que 25 membres, dont seuls 15 ont accepté de remplir les questionnaires proposés.

---

[4] http://www.wikifyer.com/

[5] https://dev.twitter.com/pages/annotations_overview (login requis)

[6] https://dev.twitter.com/discussions/3769 (login requis)

[7] http://oak.dcs.shef.ac.uk/

Deux remarques sont donc soulevées. D'une part, les échanges d'un groupe dont le point commun est les techniques d'analyse sémantiques ne sont-ils pas biaisés pour un tel exercice ? Les auteurs disent que cette liste sert autant d'échanges personnels que professionnels. Mais se servir d'un tel contexte pour valider le repérage d'experts dans toute organisation n'est-il pas potentiellement risqué ? Un éclaircissement à ce sujet aurait été le bienvenu. Par ailleurs, aucune explication n'est donnée quant au refus de 40% des membres de participer à l'expérimentation.

Poussons un peu l'analyse des résultats en gardant à l'esprit ce nombre de 15 participants. Ainsi, les auteurs balaient trois utilisateurs ne validant pas leur hypothèse d'amélioration des résultats avec la hausse du degré sémantique (IDs 18, 11 et 23 dans **Fig. 7**) en soulignant que l'**Agr** (*inter-annotator agreement*, dont les limitations de calcul ont été largement soulignées plus haut) est plus faible de 5 points chez ces utilisateurs que chez les autres. Cela suffit-il réellement à éviter la prise en compte de 20% d'utilisateurs réfutant l'hypothèse ? Il n'est pas non plus expliqué pourquoi certaines significances atteignent des niveaux ridicules[8] (0,53 ; 0,48 ; 0,32…) ; en réalité, moins de la moitié des corrélations (7 sur 15) ont une significance inférieure à 0,05, limite supérieure arbitraire généralement acceptée et pourtant déjà contestée car laissant trop de place au hasard [10, 11]. Cette insignifiance concerne les valeurs d'analyse par mots-clés, mais ne s'améliore que peu pour l'analyse par concepts (8 sur 15), et empire pour l'analyse par entités (6 sur 15). Revenons d'ailleurs sur cette analyse par entités : les auteurs ne donnent aucune explication pour la baisse — certes mineure — de corrélation (de 37% en moyenne à 36%) lors de l'augmentation du niveau de sémanticité depuis les mots-clés vers les entités nommées, préférant insister sur la hausse, là encore mineure, observée lors du passage aux concepts (43%). Mais au-delà de tous les biais soulignés jusqu'ici, cette corrélation finale inférieure à 50% n'est-elle pas à elle seule trop faible pour valider l'hypothèse ?

Sans conclusion explicite quant à leurs résultats, les auteurs enchaînent immédiatement sur une présentation des différentes visualisation auxquelles leurs données les ont menés. On admire effectivement les graphes, mais on reste surpris par les incohérences de description rendant toute compréhension des résultats impossible.

[8] Une significance *p* représente la probabilité qu'une corrélation ait été le résultat du hasard. 0 est donc excellent, et 1 totalement insignifiant puisqu'également attribuable à la chance.

Figure 8. Diagramme dirigé par les forces représentant l'expertise du groupe

La lecture donnée par les auteurs de la **Fig. 8** est la suivante[9] :
*"The concepts **closer to the central node [are] shared by the majority of users**, while nodes on the outer circle (randomly picked) are concepts shared by a small number of people […]. Figure [8] clearly highlights the **emerging topics in the group as {Semantics, Wiki, Emoticon, Technology,** […] }, but also allows to identify **topics that are emerging or have less wide-spread** […] **{Industry, Semantics, Semantic similarity, Javascript, Debate, Vegetarianism,** […] **Shoe, Chocolate}**."*

Les incohérences graves sont les suivantes : (i) Ce commentaire oppose donc des concepts "émergents" à… des "concepts émergents" (*sic*). Aucune autre catégorie n'est donnée dans l'article. (ii) Certains éléments sont dans les deux groupes ("Semantics" par exemple). (iii) Certains éléments n'apparaissent pas sur le diagramme ("Chocolate"). (iv) Soi-disant, les concepts les plus proches du nœud central sont ceux les plus communs. Il paraîtrait bien surprenant que sur une mailing-list d'académiques en gestion de la connaissance, les thèmes "Vegetarianism" ou "Shoe" soient plus partagés que "Semantic Web" ! Il suffit de toute façon de comparer ce graphe au nuage de mots-clés ensuite donné par les auteurs.

Figure 9. Nuage de mots-clés représentant l'expertise du groupe

Cette visualisation est clairement incompatible avec la lecture du graphe précédent. On pourrait donc supposer que les auteurs ont inversé la signification de la proximité dans la **Fig. 8**, mais cela n'explique pas pour autant les autres incohérences soulevées plus haut, ni pourquoi le nuage de mots-clés semble ordonner son contenu par taille et par sens de lecture avant de normaliser la taille (à partir de "Hyperlink"). D'autres

[9] c'est moi qui souligne

incohérences mineurs, telles que des disparités dans la spécificité temporelle des différents types de diagrammes, sont également rencontrées.

5. Conclusion

Ces deux articles, bien qu'utilisant tous deux l'extraction de données sémantiques depuis des échanges informels parfois différents de la langue naturelle, comme certains tweets, offrent des conclusions bien distinctes. L'un nous prouve l'efficacité de l'augmentation sémantique du contenu pour améliorer très sensiblement la recherche parmi une très grande quantité d'information ; l'autre échoue à démontrer l'intérêt de cette même technique pour repérer des liens entre personnes spécifiques et domaines de compétences. Néanmoins, la possibilité de raffiner l'algorithme de constitution du profil utilisateur dans le second article peut être envisagée. Il serait également intéressant de déterminer si la méthode de recherche validée dans le premier article pourrait être utilisée pour obtenir les résultats souhaités par le second, à savoir isoler des experts dans certains domaines au sein d'une organisation : la recherche sur certains sujets précis pourrait ainsi permettre de repérer les auteurs de messages particulièrement pertinents.

Références

1. F. Abel, I. Celik, G.J. Houben, P. Siehndel (2011) : Leveraging the Semantics of Tweets for Adaptive Faceted Search on Twitter.
2. A.L. Gentile, V. Lanfranchi, S. Mazumdar, F. Ciravegna (2011) : Extracting Semantic User Networks From Informal Communication Exchanges.
3. J. Teevan, D. Ramage, M.R. Morris (2011) : #twittersearch: a comparison of microblog search and web search. (pp. 35–44)
4. J. Diesner, T.L. Frantz, K.M. Carley (2005) : Communication networks from the Enron email: “It’s always about the people. Enron is no different”. (pp. 201–228)
5. F. Abel, Q. Gao, G.J. Houben, K. Tao (2011) : Semantic enrichment of Twitter posts for user profile construction on the social web.
6. R. Snow, B. O’Connor, D. Jurafsky, Andrew Y. Ng (2008) : Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks (p. 3)
7. R.J. Hunt (1986) : Percent Agreement, Pearson's Correlation, and Kappa as Measures of Inter-examiner Reliability (p.128)
8. J.L. Fleiss (1971) : Measuring nominal scale agreement among many raters
9. D. Milne, I.H. Witten (2008) : Learning to Link with Wikipedia
10. G.E. Dallal (2003) : Why P=0.05? *(non revu par les pairs, mais résume des points exprimés dans d'autres articles — voir les références de l'article en question, ainsi que [11])*
11. J.M. Bland, D.G. Altman (1995) : Multiple significance tests: the Bonferroni method

Bibliographie

1. Formal definition of a facet
   http://www.ai.sri.com/~okbc/okbc-faq/Knowledge_Models/what_is_facet.htm
2. How to read Pearson's Correlation Coefficient's significance
   http://lilt.ilstu.edu/gmklass/pos138/assignments/level__r.html
   http://faculty.vassar.edu/lowry/ch4apx.html
3. How is F-measure used to evaluate effectiveness of information retrieval
   Steven M. Beitzel (PhD th., 2006) : On Understanding and Classifying Web Queries (p.70)
4. Cohen's Kappa
   http://web.njit.edu/~mendonca/statcomp/papers/Cohen_nominal_scales_1960.pdf

Crédits