Atelier : veille numérique, regards pluridisciplinaires

Francis Chateauraynaud*, Jean-Gabriel Ganascia**, Julien Velcin***

*GSPR, EHESS Paris
chateau@msh-paris.fr
**LIP6, Université de Paris 6
Jean-Gabriel.Ganascia@lip6.fr
***ERIC, Université de Lyon 2
Julien.Velcin@univ-lyon2.fr

Objectifs de l'atelier

Avec les technologies de l'information, un nombre considérable de données est désormais accessible en ligne. Ces données évoluent au fil du temps et cette dynamique est riche d'enseignements. Il nous faut désormais fabriquer des outils qui aideront à en tirer parti et à lancer des alertes. C'est l'objet de la veille numérique. Cet atelier cherche plus particulièrement à engager une discussion entre les informaticiens, qui développent des approches automatiques de type statistique ou symbolique, les spécialistes de fouille de données (textuelles ou autres) et de traitement automatique de la langue, et les chercheurs en Sciences Humaines et Sociales travaillant à partir d'écrits (analyse du discours, étude des controverses, etc.). Il devra permettre de faire le point sur l'avancée des recherches réalisées par les différentes communautés sur ce domaine et sur les travaux communs qui pourraient être mis en œuvre.

Présentation de la veille numérique

De nombreux sites internet proposent aujourd'hui de consulter des brèves concernant l'actualité générale (dépêches d'agences) ou des sujets plus spécialisés (résumés d'articles scientifiques, information sur l'avancement technologique, études économiques ou politiques, etc.). Ce sont des textes courts (quelques paragraphes) rédigés en langage naturel. Ce précieux gisement est encore loin d'être pleinement exploité. Pourtant, il est susceptible d'intéresser de nombreux acteurs sociaux, que ce soient des décideurs, des assureurs, des hommes politiques, ou des scientifiques, par exemple des historiens, des sociologues, etc. Plus généralement, la veille technologique et l'intelligence économique prennent un nouvel élan du fait de la numérisation des contenus et de leur accès libre sur la toile. Nous assistons à l'essor d'un champ novateur qui aborde la veille sur les données numériques en ligne en procédant à des études diachroniques sur les contenus. Nous souhaitons que ce premier atelier soit un endroit d'échange entre plusieurs communautés abordant ces questions de veille numérique – c'est-à-dire de veille technologique, stratégique ou économique, sur des données numériques – avec des approches diachroniques.

La cartographie des brevets dans l'industrie et la recherche : outils et pratiques

Antoine Blanchard*

*Deuxième labo, 56 bd Auguste Blanqui, 75013 Paris
antoine.blanchard@gmail.com
http://www.deuxieme-labo.fr

Résumé. L'information brevet est souvent perçue comme une mine d'or pour l'industrie et comme un puits sans fond pour la recherche académique. La demande de l'industrie pour des outils lui permettant d'exploiter cette mine d'or en ont fait un utilisateur routinier de nombreuses technologies de cartographie de l'information, de fouille de texte ou de données, quand les chercheurs se contentent souvent d'une information brevet géographiquement limitée et des méthodes d'analyse qui leur sont familières. Malgré leurs différences d'approches et d'objectifs, les deux domaines peuvent apprendre l'un de l'autre. Dans cette communication, nous tenterons de rendre compte, de manière aussi exhaustive que possible, des pratiques qui ont cours dans l'industrie ou dans la recherche et de montrer comment certains dispositifs de navigation entre niveaux méso- et macroscopique pourraient les rapprocher.

1 L'information brevet et ses différents niveaux d'analyse

L'information brevet est souvent la plus redoutée des non-spécialistes, et même des spécialistes, car elle présente plusieurs difficultés :

- elle est disparate : chaque pays possède son propre système de brevets et donc ses propres brevets (les brevets relatifs à une même invention déposés dans différents pays formant ce que l'on nomme une famille de brevets, *cf.* figure 1), publiés dans sa propre langue et son propre format (base de données électronique ou non, résumés ou plein texte etc.) ;
- elle est protéiforme : des figures détaillées côtoient un vocabulaire technico-juridique ou «*patentese*» dont le but est souvent d'être le plus englobant et le moins explicite possible ;
- elle est riche de méta-données : chaque brevet va de pair avec un inventeur et un déposant mais aussi une ou plusieurs classes de la Classification internationale des brevets (CIB), un statut juridique[1], une place dans l'histoire de sa famille[2]…

À l'inverse, les publications scientifiques sont un matériau d'acccès plus facile et immédiat et donc plus couramment utilisé dans les recherches sur la sociologie et l'histoire des sciences et techniques. Pourtant, l'information brevet est irremplaçable pour comprendre notamment les processus d'innovation. Les entreprises de R&D la manient tous les jours pour des raisons évidentes et se sont constitué une expertise qui peut être valorisée dans d'autres contextes. Cette expertise relève de deux approches distinctes :

- chercher et utiliser l'information brevet avec l'œil expert de l'ingénieur brevets ou du scientifique qui s'intéresse à sa portée juridique ou son contenu technique (analyse microscopique) ;
- chercher et utiliser l'information brevet selon l'approche plus holistique du veilleur qui s'intéresse aux motifs d'ensemble, aux signaux faibles etc. (analyse macroscopique). Anthony J. Trippe (2002) a proposé de nommer «*patinformatics*» cette deuxième approche, sur le modèle de disciplines ayant récemment bouleversé la méthode scientifique par leur utilisation de données massives comme la bio-informatique ou la chemo-informatique.

Les outils de la *patinformatics*, dont la cartographie des brevets, sont venus soutenir en routine l'approche macroscopique du veilleur ou du consultant en innovation mais sont aussi déployés aujourd'hui pour l'analyse mésoscopique par les sociologues ou chercheurs en économie, gestion et propriété intellectuelle.

[1] Ainsi, un brevet dont la redevance cesse d'être payée perd son effet légal au bout de plusieurs mois.

[2] On distingue en particulier entre la demande de brevet, qui n'a aucun effet légal, et le brevet proprement dit qui est délivré après examen et peut différer substantiellement de la demande de brevet. Les États-Unis sont aussi fameux pour leur système complexe qui donne naissance à des familles de brevets arborescentes plutôt que linéaires.

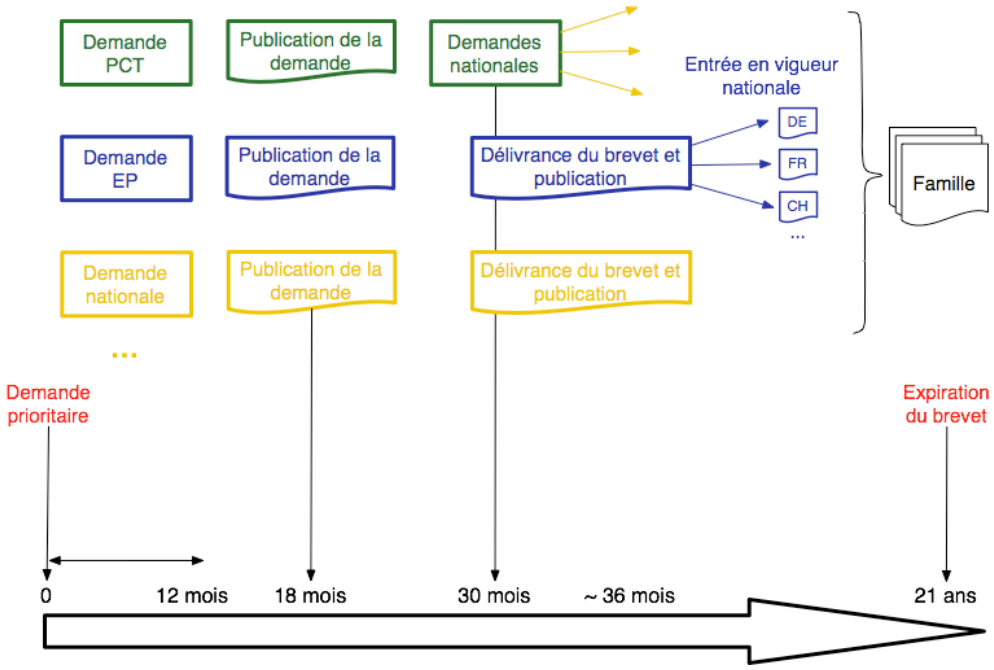


FIG. 1 – *Cheminement typique d'une demande de brevet résultant en une famille de brevets, depuis le dépôt au niveau national ou international («EP» pour le système européen, «PCT» pour le système de l'Organisation mondiale de la propriété intellectuelle…) dans les 12 mois suivant la demande prioritaire jusqu'à la délivrance d'une famille de brevets et leur expiration.*

2 Différents outils et pratiques de cartographie des brevets

La cartographie des brevets est définie par l'Office européen des brevets comme la représentation visuelle du résultat d'analyses statistiques ou de fouille de texte appliquées aux brevets, permettant de comprendre et évaluer facilement de larges volumes d'information brevet[3]. Pour Anthony J. Trippe (2002), cela revient essentiellement à représenter graphiquement ou physiquement dans un espace-plan l'art relevant d'un domaine particulier. Techniquement, la cartographie regroupe l'ensemble des représentations graphiques de données groupées, allant notamment de l'analyse dimensionnelle comme l'analyse en composantes principales à la classification non-supervisée en passant, dans une moindre mesure, par la construction d'arbres ou dendrogrammes (Porter et Cunningham, 2005, chapitre 10.3).

Parmi ses traits distinctifs, on trouve en particulier la présence de regroupements appelés «*clusters*», qu'ils soient des partitions définies rigoureusement ou des représentations graphiques lâches, séparés par une certaine distance ou des arêtes. Un *cluster* peut être obtenu de façons très diverses et ne se limite pas ici à la classification non-supervisée comme le voudrait sa définition stricte.

La cartographie des brevets recouvre donc une palette d'outils et de techniques, complémentaires, qui peuvent être utilisés aussi bien par l'industrie que par la recherche. Nous allons les détailler à présent, en gardant à l'esprit que chaque technique peut être déclinée selon une typologie transversale, celle de la dimension temporelle :

- la cartographie statique donne une image figée à un certain point dans le temps, qui est généralement le moment de l'analyse mais représente souvent le passé en raison des délais propres à l'information brevet (18 mois entre le dépôt de la demande prioritaire et la première publication, plus le temps d'enregistrement et d'indexation dans les bases de données) ;
- la cartographie dynamique peut être utilisée pour retracer le développement au cours du temps d'une technologie. Elle consiste souvent en une suite de cartes statiques, qui découpent la période temporelle étudiée.

2.1 Les cartes sémantiques ou «*topic maps*»

L'approche des cartes sémantiques consiste à laisser parler le corpus par lui-même en se reposant uniquement sur son contenu lexical, plutôt que sur des catégories définies *a priori* (classifications des bases de données bibliographiques,

[3] http://www.epo.org/patents/patent-information/business/stats/faq.html#mapping

mots-clés attribués par les auteurs) ou *a posteriori* (jugement des experts). Cette approche permet de se débarasser des idées préconçues et donc de mieux rendre compte du front de recherche, en perpétuel renouvellement (Mogoutov et al., 2008). Elle permet aussi d'ajuster le degré de détail au niveau voulu, comme nous le montrerons dans la partie 3.

Méthodologie. Il faut en principe deux étapes pour obtenir une carte sémantique. La première découpe le corpus étudié en *n* entités (termes et expressions) bien délimitées tandis que la seconde représente le résultat obtenu sous la forme de carte, typiquement en projetant en deux dimensions l'espace de vecteurs (représentant chacun un document) à *n* dimensions. Ainsi, plus le contenu lexical des documents est proche, plus la distance qui les sépare est faible (les deux axes de la carte n'ont aucune signification en soi), la densité des documents pouvant être représentée en sus par des « montagnes » (Blanchard, 2006). Certains logiciels « clés en mains » combinent les deux étapes, qui deviennent indiscernables pour l'utilisateur.

Trousse à outils. Deux familles de méthodes permettent la première étape de découpage du corpus, même si les frontières tendent à se brouiller avec le développement de logiciels empruntant le meilleur des deux mondes :

- la fouille de texte ou « *text mining* » utilisant exclusivement une approche statistique (fréquences d'occurrence, de co-occurrence etc.), qui réussit d'autant mieux que le corpus de documents est volumineux. C'est la méthode par défaut des logiciels de cartographie clés en mains, notamment ceux qui sont spécialisés dans l'information brevet comme STN AnaVist[4] ou Aureka[5] ;

- la fouille de texte à base de traitement automatique des langues ou « *natural language processing* » (NLP) qui utilise l'analyse syntaxique des phrases, caractérise la relation entre différents termes, résout les ambiguïtés du langage et peut utiliser un dictionnaire spécialisé. Cette méthode est l'apanage de logiciels souvent professionnels, comme SPSS LexiQuest Mine[6] utilisé par Mogoutov et al. (2008), auxquels il faut adjoindre un logiciel de représentation des données (ReseauLu[7] en l'occurrence).

Pratique des cartes sémantiques. Les outils clés en main fleurissent pour la cartographie sémantique des brevets, et sont utilisés avidemment par l'industrie (Eldridge, 2006). Ils sont faciles à prendre en main, faciles à interpréter pour la prise de décision et permettent de se passer de la CIB, employée par l'ensemble des examinateurs des offices des brevets à travers le monde, qui manque souvent de pertinence pour ses besoins (Fattori et al., 2003) et de clarté pour ses utilisateurs finaux. Mais la recherche s'en est également emparée à quelques occasions, comme dans cette étude des brevets dans le domaine des biotechnologies pour l'agriculture (Graff et al., 2003). Le logiciel utilisé, Aureka, permet notamment de créer automatiquement des tranches temporelles dans les données et de comparer l'activité technologique entre plusieurs périodes, comme illustré par Trippe (2001).

Informations pouvant enrichir une carte sémantique. De nombreuses informations peuvent être ajoutées à une carte sémantique en utilisant des couleurs ou la forme des points, par exemple l'information sur le déposant du brevet ou le fait qu'il appartienne au corpus citant ou cité, pour lui faire répondre à des questions encore plus larges. Chez Syngenta Crop Protection, entreprise du secteur agrochimique, nous avons utilisé cette approche pour repérer où se concentrent les citations par rapport au portefeuille de brevets et caractériser l'activité des concurrents ou de l'industrie pharmaceutique autour de certaines classes de molécules brevetées par Syngenta.

2.2 Les cartes macro-thématiques

Ces cartes mobilisent ce que Zitt et Bassecoulard (2004) qualifient de dimension macro-thématique de l'information scientifique, c'est-à-dire aussi bien l'indexation manuelle offerte par les bases de données à valeur ajoutée que les diverses classifications appliquées aux documents. La CIB a ainsi été utilisée par Meyer (2007) dans une étude de cas sur les nanotechnologies et les codes de la base de données Derwent[8] par Trippe (2001) dans l'étude du portefeuille de l'entreprise pharmaceutique Vertex.

[4] http://www.stn-international.de/stninterfaces/stnanavist/stn_anavist.html
[5] http://www.thomsonreuters.com/products_services/scientific/Aureka
[6] http://www.spss.com/fr/lexiquest/lexiquest_mine.htm
[7] http://www.aguidel.com/fr/?sid=6
[8] http://www.stn-international.de/stndatabases/databases/wpidswpx.html

Carte de co-occurrences. Meyer (2007) identifie 5407 sous-classes de la CIB présentes dans son corpus puis range par deux les sous-classes qui apparaissent ensemble sur un ou plusieurs brevets, avec leur fréquence de co-occurrence. Il utilise le logiciel de statistique SPSS[9] pour représenter cette matrice sur une carte, par une méthode de «*Multi-Dimensional Scaling*» (MDS) présente dans la routine SYSTAT de l'algorithme ALSCAL. Chaque sous-classe de la CIB apparaît d'autant plus grosse qu'elle est fréquente dans le corpus et proche des autres sous-classes avec lesquelles elle co-occurre souvent. On visualise ainsi des *clusters* de technologies composant le champ des nanotechnologies, où les dispositifs de mesure se distinguent des nano-matériaux, du monde de la chimie ou pharmacie et des nano-couches ou semi-conducteurs.

Regroupement ou «clustering» des brevets. Trippe (2001) utilise le logiciel de fouille de données TWID Expert[10] pour regrouper les brevets et demandes de brevets qui sont codés de façon similaire par les indexeurs Derwent. Pour ce faire, le motif d'ensemble des codes de classification est pris en compte («*many to many*») et pas uniquement leurs co-occurrences deux à deux. Les 128 documents ainsi traités se retrouvent regroupés dans 25 *clusters* (nombre fixé par l'analyste par essai et erreur), avant d'être rangés et étiquetés manuellement.

2.3 Les réseaux de citation

À l'inverse des méthodes présentées précédemment, les réseaux de citation reposent sur des liens directs et directionnels entre documents, à savoir les liens de citation. La cartographie consiste principalement à les représenter graphiquement et à en tirer des informations complémentaires par le jeu de la transformation et de l'analyse, comme cela se fait couramment dans l'analyse des articles scientifiques (Small, 1999).

Représentation des liens de citation. Les citations entre brevets sont un indicateur de la proximité technologique entre différentes inventions, dans certaines limites soulignées par Oppenheim (2000)[11]. En agrégeant les données par déposant et les mobilisant dans des réseaux de citations à grande échelle, on peut mettre en évidence certaines relations entre les acteurs — les plus liés entre eux étant les plus susceptibles d'entrer en compétition. En agrégeant les brevets en *clusters* ayant un profil de citation similaire, on peut aussi représenter les flux de connaissances («*knowledge flows*») à un niveau mésoscopique. C'est ainsi que Igami (2008) fait apparaître le rôle de pivot des capteurs et actionneurs pour les nanotechnologies alors que des technologies comme les cristaux photoniques ou la nanolithographie sont beaucoup plus isolées.

Le réseau peut être calculé sur la base des brevets individuels plutôt que sur celle des déposants, à condition de fixer un seuil ou une limite temporelle pour diminuer le nombre d'observations et conserver un résultat lisible. On peut ainsi être amené à s'intéresser à un brevet en particulier, comme celui de l'entreprise japonaise Rohm dont Sternitzke et al. (2008) constatent qu'il cite pas moins de 229 familles de brevets appartenant en partie au concurrent Nichia, et ce en raison de ses revendications particulièrement larges. On retombe ici dans une analyse plus microscopique et Aureka offre cette possibilité en routine, très appréciée en entreprise.

Abstraction des liens de citation. Une autre approche consiste à utiliser les liens de citation pour calculer un facteur d'attraction ou de répulsion entre chaque brevet. En utilisant un algorithme développé pour la représentation de molécules en chemo-informatique, Masatsura Igami (2008) obtient ainsi des *clusters* dont ont disparu les liens de citation à proprement parler, remplacés par une distance plus ou moins grande entre *clusters* et en leur sein. Cette étude de cas sur les nanotechnologies lui permet par exemple d'observer quinze domaines constitutifs, contenant de 43 à 550 demandes de brevets. Une série temporelle lui permet également de mettre en évidence l'évolution de ces domaines technologiques entre 1990 et 2000, notamment dans leurs interactions.

Pour aller plus loin. Les citations allant toujours du plus récent au plus ancien, elles permettent aussi de retracer la trajectoire d'un domaine et l'évolution des centres d'intérêt des inventeurs (en corrélant par exemple la taille des points du réseau à l'âge des brevets, comme chez Sternitzke et al. (2008), ou en indiquant le sens de la citation). Pour identifier selon ce principe le progrès des technologies d'actionnement variable de soupape dans l'industrie automobile, von Wartburg et al. (2005) ont constitué un corpus selon une méthode itérative qui s'appuie sur la chaîne des citations de brevets européens. Le résultat montre par exemple que l'ajout d'un arbre à cames par l'industrie est venu compléter aussi bien les solutions hydrauliques que mécaniques, lesquelles forment deux *clusters* bien distincts. Dans un second temps, les auteurs vont encore plus loin en tenant également compte du couplage bibliographique (le fait pour deux brevets de

[9] http://www.spss.com/fr/statistics/

[10] http://www.synthema.it/textmining

[11] Et il est exagéré d'en faire autre chose, comme Sun et Morris (2008) qui y voient la marque de l'utilisation d'une technologie par une autre.

citer un même document) et des co-citations (le fait pour deux brevets d'être cités par un même document), pour obtenir une carte de proximité qui confirme les observations précédentes.

Le couplage bibliographique est utilisé également par Sun et Morris (2008) pour regrouper les brevets qui ont des motifs de citation similaires, formant 10 *clusters* relatifs à la technologie des têtes de lecture optiques. Ils décomposent alors chaque *cluster* selon un axe temporel pour montrer l'évolution des fronts de recherche, dont on voit qu'ils se relaient les uns les autres. En complétant la représentation graphique avec des donnés de citation (chaque point apparaît d'autant plus gros que le brevet a été beaucoup cité et d'autant plus foncé qu'il est encore cité aujourd'hui), ils montrent en même temps lesquelles de ces lignes de recherche ont été les plus fécondes.

Il ne faut pas oublier non plus les liens d'auto-citation qui permettent de visualiser l'importance de l'innovation endogène face au transfert de technologies extérieures (Jaffe, 1998) ou l'enchevêtrement des brevets d'une entreprise se citant abondamment les uns les autres (qui forment des «*patent thickets*» défavorables aux compétiteurs qui voudraient entrer dans le domaine).

2.4 Les cartes de réseaux sociaux

Les brevets peuvent aussi être utilisés non pas pour eux-mêmes, mais pour faire parler les acteurs qui se cachent derrière, individus ou institutions. C'est ce que viennent de réaliser Sternitzke et al. (2008) dans une étude de cas portant sur les diodes électroluminescentes et les diodes lasers.

Trousse à outils. Certains logiciels tournés vers l'information brevet proposent l'analyse clés en mains des réseaux de collaboration entre inventeurs ou déposants, à l'instar de Matheo Patent[12] (Dou, 2004), de Vantage Point[13] ou de Thomson Data Analyzer[14]. Mais l'utilisation de logiciels d'analyse et de cartographie de réseaux sociaux comme ReseauLu, Pajek[15] ou UCINET[16] offre une plus grande souplesse à l'analyste, à condition qu'il prenne la peine de reformater l'information issue des bases de donnés de brevets dans des formats compatibles, à l'aide par exemple de PATONanalist[17] (Sternitzke et al., 2007).

Préparation et représentation des données. Les données sont susceptibles d'être faussées par la présence d'homonymes ou synonymes et doivent de préférence être nettoyées. Une solution proposée par Sternitzke et al. (2008) consiste à croiser les données avec la base INPADOCDB[18] de l'Office européen des brevets, laquelle contient le nom complet des inventeurs et pas seulement l'initiale de leur prénom. Selon l'analyse recherchée, les nœuds du réseau seront les brevets, les inventeurs ou les déposants, dont la loi de Lotka (1926) prédit que l'immense majorité ont en fait une contribution anecdotique. Il est donc conseillé de se limiter aux inventeurs ou déposants les plus actifs, par exemple ceux du dernier décile, ce qui permet en même temps de rendre le résultat plus lisible.

Réseaux de coopération. En première approximation, la coopération entre inventeurs ou déposants revient à la co-signature ou le co-dépôt de brevets. L'utilisation de cette information contenue dans les champs bibliographiques des brevets permet de tracer les réseaux de coopération et mettre en évidence des sous-groupes fortement connectés («collèges invisibles»), des acteurs qui sont centraux et ceux qui font office de passeurs entre différents sous-groupes. En utilisant la CIB qui décrit l'appartenance de chaque brevet à un ou plusieurs domaines technologiques, Sternitzke et al. (2008) montrent notamment que ces passeurs ont une activité significativement plus variée, car ils ont accès au savoir de plusieurs collèges invisibles.

2.5 Les cartes géographiques

L'information brevet contient deux niveaux d'information d'intérêt géographique : le pays d'appartenance du déposant (ou inventeur) et le pays de dépôt du brevet. Igami (2008) a combiné le premier niveau avec une carte de densité pour comparer la spécialisation technologique de la Suisse et de la Corée, représentant chacune un sous-ensemble du corpus sur les nanotechnologies. Le deuxième niveau censé montrer quels pays sont incontournables pour un domaine donné apporte en fait une information moindre puisqu'on retrouve généralement les États-Unis, le Japon et les principaux pays européens en tête de liste.

[12] http://www.matheo-patent.com
[13] http://www.thevantagepoint.com
[14] http://www.thomsonreuters.com/products_services/scientific/Thomson_Data_Analyzer
[15] http://pajek.imfm.si
[16] http://www.analytictech.com/ucinet/ucinet.htm
[17] http://www.paton.de
[18] http://www.stn-international.de/stndatabases/databases/inpadocdb.html

3 Naviguer entre niveaux méso et macro

Nous avons indiqué précédemment que l'industrie se caractérise par son besoin à la fois d'information brevet microscopique guidant les actions légales (dépôt, opposition, plainte pour infraction…) et d'analyse macroscopique orientant l'innovation, tandis que la recherche s'intéresse surtout à l'information de niveau mésoscopique. Concrètement, cela signifie que l'industrie a les moyens de cartographier régulièrement le portefeuille complet d'une ou plusieurs multinationales ou d'un large domaine technologique en multipliant les sources d'information[19] tandis que la recherche académique s'intéresse souvent à un domaine plus restreint tel que représenté dans une unique juridiction (brevets américains, européens ou français) : elle peut se contenter d'un échantillon de l'information existante quand l'industrie a besoin d'exhaustivité. Alors que l'industrie cherche à agir, la recherche vise surtout à s'informer et les enjeux ne sont évidemment pas les mêmes.

Pourtant, ces deux approches sont loin d'être antinomiques. Nous avons montré comment les mêmes outils peuvent être utilisés par les uns ou les autres pour poser des questions semblables à un corpus plus ou moins large. Mais pour un corpus et une cartographie donnés, la représentation et la navigation peuvent aussi être personnalisées selon les besoins. Notre expérience nous a notamment amené à expérimenter l'effet des listes de mots vides ou «*stopwords*» sur la granularité des cartes sémantiques, tandis que d'autres auteurs ont misé sur le choix du système de pondération ou la reconnaissance des expressions par NLP.

3.1 Personnalisation des mots vides

Les mots vides sont nés dans les bases de données des années 1950, quand les ressources informatiques limitées nécessitaient que tous les mots ne soient pas indexés comme mots clés, particulièrement ces mots grammaticaux qui représentent jusqu'à la moitié des occurrences («le», «sur», «tant»…). Depuis, ils ont quasiment disparu des bases de données de brevets mais prospèrent dans les outils de cartographie sémantique, où ils contribuent grandement à la qualité et la pertinence du résultat final (Blanchard, 2007). Ces outils sont ainsi fournis avec une liste de mots vides, également appelé antidictionnaire, couvrant au moins les besoins de la langue anglaise. Mais ces listes par défaut manquent du vocabulaire propre aux brevets avec des termes juridiques comme «*embodiment*» ou «*comprising*» et des termes lexicalement pauvres comme «*exhibit*» ou «*demonstrate*». Ceux-ci peuvent être ajoutés manuellement dans la plupart des cas.

Mais l'analyste peut aller plus loin en ajoutant aux mots vides des termes spécifiques au domaine étudié qui n'apportent pas d'information dans le contexte. Par exemple, le terme «*protéine*» est superflu dans l'analyse de l'utilisation d'enzyme phytase pour la nutrition animale, dans la mesure où toutes les enzymes sont des protéines. L'expérience montre que cette méthode permet d'obtenir des *clusters* mieux séparés et d'affiner le niveau de l'analyse autour de concepts plus discriminants (Blanchard, 2007; Trippe, 2001), «moins macroscopiques». Elle peut même être automatisée chez certains outils comme OmniViz[20] et STN AnaVist.

3.2 Choix du système de pondération

Le système de pondération ou «*weighting system*» est un paramètre de l'algorithme de fouille de texte qui permet d'obtenir soit des gros *clusters* reposant sur des mots qui reviennent fréquemment, soit des petits *clusters* reposant sur des mots plus rares (Fattori et al., 2003). Un système fréquemment utilisé est l'algorithme «tf $\times$ idf» (pour «term frequency $\times$ inverse document frequency»), qui consiste à faire le produit de deux mesures complémentaires : la fréquence du mot dans le corpus et l'inverse de la proportion de documents qui contiennent ce mot (Salton et Buckley, 1988). Ainsi, le poids d'un mot augmente quand il est utilisé abondamment et il diminue quand il apparaît dans de nombreux documents. La personnalisation du système de pondération est permise par quelques outils comme PackMOLE, développé en interne pour les besoins de l'entreprise Tetra Pak (Fattori et al., 2003). Selon l'expérience de ces auteurs recherchant à obtenir des *clusters* à la fois homogènes et bien séparés, l'option de pondération préférée est celle favorisant des *clusters* spécialisés à base de mots plus rares.

3.3 Reconnaissance des expressions

Mogoutov et al. (2008) indiquent que le NLP permet de gagner un niveau de pertinence dans l'analyse par rapport à la fouille de texte statistique, en facilitant le travail préalable à la cartographie par la reconnaissance des expressions. Dans leur exemple, plutôt que de s'embarrasser d'un mot comme «cancer» qui est tellement courant dans leur corpus sur les puces à ADN qu'il en devient trivial, l'algorithme isole des expressions assez fréquentes comme «cancer du sein», «cancer du côlon» ou «épidémiologie du cancer». Ainsi, au lieu d'attendre que la cartographie fasse ressortir des *clusters*

[19] C'est ainsi que chez Syngenta, nous avons cartographié le portefeuille actif de l'entreprise, soit plus de 1 000 familles de brevets couvrant trois juridictions de délivrance de brevets.

[20] http://www.omniviz.com/content/omniviz

significatifs, on affine le découpage du corpus en amont et on gagne en finesse. À condition que cette finesse ne devienne pas excessive, auquel cas on peut toujours ajouter les nouvelles expressions jugées superflues à la liste de mots vides, selon le modèle précédent.

4 Conclusion

La cartographie des brevets désigne un ensemble varié de pratiques visant à représenter graphiquement un corpus de brevets et les informations issues de son analyse. Elle possède un fort pouvoir de séduction visuelle dans l'industrie et la recherche académique est convaincue de son intérêt heuristique depuis au moins les premières propositions de Derek J. De Solla Price (1965) visant à cartographier les articles scientifiques. On peut néanmoins ranger les outils et pratiques énumérés ici en plusieurs familles, selon qu'ils s'intéressent aux brevets (exemple des cartes sémantiques) ou aux acteurs (réseaux sociaux), aux données primaires (origine géographique) ou secondaires (réseaux de citation), aux données structurées (classifications macro-thématiques) ou non-structurées (cartes sémantiques). Certaines privilégient l'analyse quantitative des données quand d'autres cherchent à découvrir des motifs invisibles à première vue ou encore à mesurer la valeur qualitative d'un portefeuille de brevets grâce à l'analyse de citations.

Malgré tous ses attraits et retombées potentielles, il ne faut pas oublier que la cartographie des brevets est contrainte par les difficultés que nous soulignions dans le premier chapitre. Si les données sont inadaptées ou mal comprises, l'analyse ne vaut rien du tout. L'expérience montre que l'industrie tombe souvent dans ce piège, notamment à cause de la tentation du « tout, tout de suite », alors que la recherche est plus habituée à la rigueur et au doute systématique. La normalisation (Porter et Cunningham, 2005, chapitre 12.6), par exemple, ou la comparaison avec des études similaires effectuées dans d'autres domaines ou des études différentes effectuées dans le même domaine devraient faire partie des réflexes indispensables au spécialiste de la cartographie des brevets. Quant à la question de la reproductibilité de l'analyse, elle est résolue par certains auteurs comme Mogoutov et al. (2008) en utilisant des logiciels standards du commerce.

Références

Blanchard, A. (2006). La cartographie des brevets. *La Recherche 398*, 82–83.

Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information 29*(4), 308–316.

De Solla Price, D. J. (1965). Networks of scientific papers. *Science 149*(3683), 510–515.

Dou, H. J.-M. (2004). Benchmarking R&D and companies through patent analysis using free databases and special software : A tool to improve innovative thinking. *World Patent Information 26*(4), 297–309.

Eldridge, J. (2006). Data visualisation tools–a perspective from the pharmaceutical industry. *World Patent Information 28*(1), 43–49.

Fattori, M., G. Pedrazzi, et R. Turra (2003). Text mining applied to patent mapping : A practical business case. *World Patent Information 25*(4), 335–342.

Graff, G. D., S. E. Cullen, K. J. Bradford, D. Zilberman, et A. B. Benett (2003). The public-private structure of intellectual property ownership in agricultural biotechnology. *Nature Biotechnology 21*(9), 989–995.

Igami, M. (2008). Exploration of the evolution of nanotechnology via mapping of patent applications. *Scientometrics 77*(2), 153–171.

Jaffe, A. B. (1998). Patents, patent citations, and the dynamics of technological change. Technical report, National Bureau of Economic Research.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16*(12), 317–323.

Meyer, M. (2007). What do we know about innovation in nanotechnology ? Some propositions about an emerging field between hype and path-dependency. *Scientometrics 70*(3), 779–810.

Mogoutov, A., A. Cambrosio, P. Keating, et P. Mustar (2008). Biomedical innovation at the laboratory, clinical and commercial interface : A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics 2*(4), 341–353.

Oppenheim, C. (2000). Do patent citations count ? In B. Cronin et H. Barsky Atkins (Eds.), *The web of knowledge : A festschrift in honor of Eugene Garfield*, ASIS&T Monograph Series, pp. 405–432. Medford, NJ : Information Today Inc.

Porter, A. L. et S. W. Cunningham (2005). *Tech mining : Exploiting new technologies for competitive advantage*. New York : Wiley InterScience.

Salton, G. et C. Buckley (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management 24*(5), 513–523.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science 50*(9), 799–813.

Sternitzke, C., A. Bartkowski, et R. Schramm (2008). Visualizing patent statistics by means of social network analysis tools. *World Patent Information 30*(2), 115–131.

Sternitzke, C., A. Bartkowski, H. Schwanbeck, et R. Schramm (2007). Patent and literature statistics — The case of optoelectronics. *World Patent Information 29*(4), 327–338.

Sun, T. et S. A. Morris (2008). Timeline and crossmap visualization of patents. In H. Kretschmer et F. Havemann (Eds.), *Proceedings of WIS 2008*, Humboldt Universität, Berlin, Allemagne.

Trippe, A. J. (2001). A comparison of ideologies : Intellectually assigned co-coding clustering vs ThemeScape automatic thememe mapping. In *Proceedings of the 2001 Chemical Information Conference*, Nîmes, France, pp. 61–79.

Trippe, A. J. (2002). Patinformatics : Identifiying haystacks from space. *Searcher 10*(9), 28.

von Wartburg, I., T. Teichert, et K. Rost (2005). Inventive progress measured by multi-stage patent citation analysis. *Research Policy 34*(10), 1591–1607.

Zitt, M. et E. Bassecoulard (2004). S&T networks and bibliometrics : The case of international scientific collaboration. In *4th Congress on Proximity Economics : Proximity, Networks and Co-ordination*, Marseille, France. GREQAM – IDEP.

Summary

The industry resorts to numerous information mapping, text mining and data mining technologies to tap into the patent information goldmine while academic researchers often content themselves with geographically limited information and familiar analysis tools. In this communication, we shall review as comprehensively as possible the practice of both domains and show how they could be brought closer to each other.

Veille sociologique et flux d'informations numériques

Une expérience de chronique automatique sur les thèmes sanitaires et environnementaux

Francis Chateauraynaud et Josquin Debaz
GSPR (EHESS)
chateau@msh-paris.fr debaz@ehess.fr

Texte pour l'atelier
« veille numérique, au carrefour des sciences »
EGC 2009

Décembre 2008

Depuis plus d'une dizaine d'années, les travaux de sociologie sur les alertes et les controverses publiques se sont accompagnés de développements informatiques spécifiques. Le principal objectif de ces travaux est d'outiller les enquêtes, en permettant de lier le suivi de grands dossiers, notamment des dossiers sanitaires, environnementaux ou technologiques (amiante, nucléaire, OGM, nanotechnologies, pesticides, champs électro-magnétiques), et la modélisation des formes de l'action et du jugement utilisées par de multiples protagonistes. Il ne s'agit pas de développer une sociologie des discours ou des textes mais bien de considérer les ensembles discursifs comme des lieux de cristallisation de processus sociaux complexes marqués par une profonde réflexivité des acteurs, puisque ceux qui s'imposent sont aussi ceux qui concentrent le plus d'expertise – ou de contre-expertise – sur les dossiers en question[1]. Dans le cadre de ces travaux, pour faire face à l'accumulation documentaire et aux changements de phases, comme lorsque se produit un emballement politique ou médiatique sur un sujet, il est décisif de disposer de protocoles de « veille ». La fonction de ces protocoles est d'assurer dans le même mouvement la mise à jour des bases documentaires et l'identification des changements de régimes argumentatifs, l'émergence de nouveaux acteurs ou d'événements et de prises de parole susceptibles de modifier les éléments centraux d'un dossier.

A travers la description d'un prototype de chroniqueur automatique, ce texte aborde différents problèmes posés par l'automatisation d'une veille sociologique sur des séries textuelles évolutives : quelles sources utiliser ? Comment éviter à la fois dispersion et redondance ? Quels outils statistiques et sémantiques mettre en oeuvre pour faire parler les séries collectées ? A quelles séries comparables, quels fonds documentaires ou quelles encyclopédies de connaissances doit-on rapporter les documents analysés pour caractériser leur contenu et surtout leur apport du point de vue pragmatique ? Quelles opérations évaluatives mettre en place et quelle place accorder à l'utilisateur, et donc à l'interprète dans la boucle ? Une veille non supervisée a-t-elle un sens ? Le produit d'un outil de veille est-il équivalent à un agrégateur sémantique ? Enfin, quel mode d'archivage et de réinterrogation doit-on se donner ? Vastes questions ! Dans cette courte contribution, nous

[1] F. Chateauraynaud, Prospéro : Une technologie littéraire pour les sciences humaines, Paris, CNRS, 2003 ; F. Chateauraynaud, « *Moteurs de (la) recherche et pragmatique de l'enquête. Les sciences sociales face au Web connexionniste* », in L'Historien face à l'ordre informatique, Matériaux pour l'histoire de notre temps, n°82, avril-juin 2006, (revue de la BDIC), p. 109-118.

proposons d'alimenter la discussion en examinant les pistes et les problèmes suscités par la réalisation d'un agent chroniqueur dédié au fil santé-environnement de nos travaux[2].

Suivre l'évolution de grands dossiers : entre acteurs hétérogènes et séries homogènes

Le programme socio-informatique développé depuis plusieurs années se place à la croisée de deux tendances lourdes : d'une part le déploiement de la « société de l'information » dont une des dimensions repose sur une conception coopérative et collective des artefacts cognitifs[3] ; de l'autre, les transformations massives de ce que l'on appelle la « société du risque »[4]. En moins de dix ans, la problématique des risques a changé complètement de nature et de forme. La période des crises, amorcée en France avec l'affaire du sang contaminé qui se déploie véritablement dans l'espace public en 1991, atteint une forme de point d'acmé avec l'enchaînement consécutif de quatre grands dossiers : l'amiante qui est de retour après quinze ans de « silence », le dossier nucléaire qui défraye la chronique avec des alertes à La Hague puis les dix ans de Tchernobyl (1996), la crise de la vache folle puis l'arrivée des premiers OGM en France – arrivée dénoncée par Greenpeace qui intercepte une cargaison en provenance des États-Unis à l'automne 1996. Tous ces dossiers ont un point commun : les sciences et les techniques y jouent un rôle décisif, et se pose de manière cruciale la question de la validité des expertises. C'est aussi au cours de l'année 1996 que se généralisent les références au « principe de précaution » déjà issu d'une longue série de travaux et de discussions, et en particulier le Sommet de Rio de 1992. Au tournant du siècle, les sources d'alerte et les événements dramatiques saturent les arènes politico-médiatiques, avec en outre une extension de la question des risques au terrorisme sous toutes ses formes. On enregistre, dans les interventions publiques, la présence de plus en plus forte des opérations de mise en série, pointant vers l'idée que l'on est entré dans une ère d'instabilité et d'incertitude accrues. Selon les contextes, se trouvent associés, selon des figures variables, des événements comme l'incendie du tunnel du Mont-Blanc (1999), l'explosion de l'usine d'AZF (septembre 2001), le crash du Concorde (juillet 2000), la crise de la vache folle (1996-2000), celle du SRAS (2003) puis de la grippe aviaire (2005-2007), la montée de l'alerte globale sur le climat et l'occurrence du « big one » avec le tsunami de décembre 2004, ou dans un autre registre les attentats du 11 septembre 2001 puis ceux de Madrid (mars 2004) ou de Londres (juillet 2005). Ici ou là, des ponts s'effondrent, des avions s'écrasent, des déchets toxiques se répandent, des ferries se renversent, des incendies font rage, des inondations dévastent des régions entières. Même ce qui est destiné à guérir, à traiter ou à remplacer (médicaments, produits de substitution, innovations technologiques, dont les fameux nanomatériaux) engendre méfiance et inquiétude, alerte et polémique. On note ainsi de nombreuses affaires de « retrait de produit », qu'il s'agisse de cocktails thérapeutiques ou de jouets pour enfants. Aucun milieu de vie et domaine d'activité ne semble épargné, de sorte que tout signe précurseur qui parvient à un degré suffisant de visibilité publique produit une vive agitation qui engendre à peu près toujours la même configuration : des acteurs annoncent l'imminence de la catastrophe, les journalistes convoquent des experts qui, généralement, ne sont pas d'accord, les pouvoirs publics et les industriels se déclarent vigilants et mettent en place des comités ou des commissions permettant de juguler le danger. C'est dans l'organisation collective de ces dispositifs que la problématique de la « veille » et des « signaux faibles » est convoquée[5].

Dans cette nouvelle configuration, la contribution de l'internet à la prolifération des signes, des annonces, des événements, des débats publics et des mobilisations est massive. Au silence et aux rapports de pouvoirs hiérarchiques de la période antérieure s'oppose une problématique de la prolifération et de l'hétérogénéité des jeux d'acteurs et d'arguments dont les réseaux s'entremêlent à l'infini. Pour s'affranchir des effets

---

[2] Une première description de ce chroniqueur est fournie dans A. Bertrand, F. Chateauraynaud et D. Torny, Expérimentation d'un observatoire informatisé de veille sociologique à partir du cas des pesticides, rapport de la convention GSPR / AFSSET, octobre 2007.

[3] Voir G. Bowker, L. Star, W.A Turner, L. Gasser (eds.), *Social Science, Technical Systems and Cooperative Work: Beyond the Great Divide*, New Jersey: Lawrence Erlbaum Associates, 1997 ; T. Malsch, "Naming the Unnamable: Socionics or the Sociological Turn of/to Distributed Artificial Intelligence" draft sent by the author, september 1999.

[4] B. Adam, U.Beck and J.Van Loon (eds), The Risk Society and Beyond, Sage, 2000.

[5] Voir sur ce point F. Chateauraynaud, «Visionnaires à rebours. Des signaux faibles à la convergence de séries invisibles », Document du GSPR, décembre 2007. En ligne sur http://gspr.ext.free.fr/

produits par l'évolution continue des nœuds et des liens, il faut se doter d'*outils alternatifs*, capables de ramener les séries pertinentes dans un laboratoire afin de les faire parler en rompant le cycle infernal de la navigation sans fin. Il existe toutes sortes d'outils pour traquer et tracer les sources et les liens, et notamment les outils de webcrawling[6]. Mais l'ordre cartographique qui tend à s'imposer avec le web, tout en fournissant de nouveaux outils de totalisation, nous éloigne de la saisie des récits et des arguments qui est au cœur des démarches d'enquête et d'érudition en sciences humaines et sociales, et des efforts de modélisation qui permettent de doter les outils d'un minimum de cohérence et de robustesse. En l'absence d'outil d'évaluation des contextes et des niveaux d'information, comment décider du choix des sites ou des blogs à visiter en priorité ? En fonction de leur position relative dans le jeu des intercitations ou des interconnexions ? On retrouve ici les critiques faites à Google et au système de classement des pages.

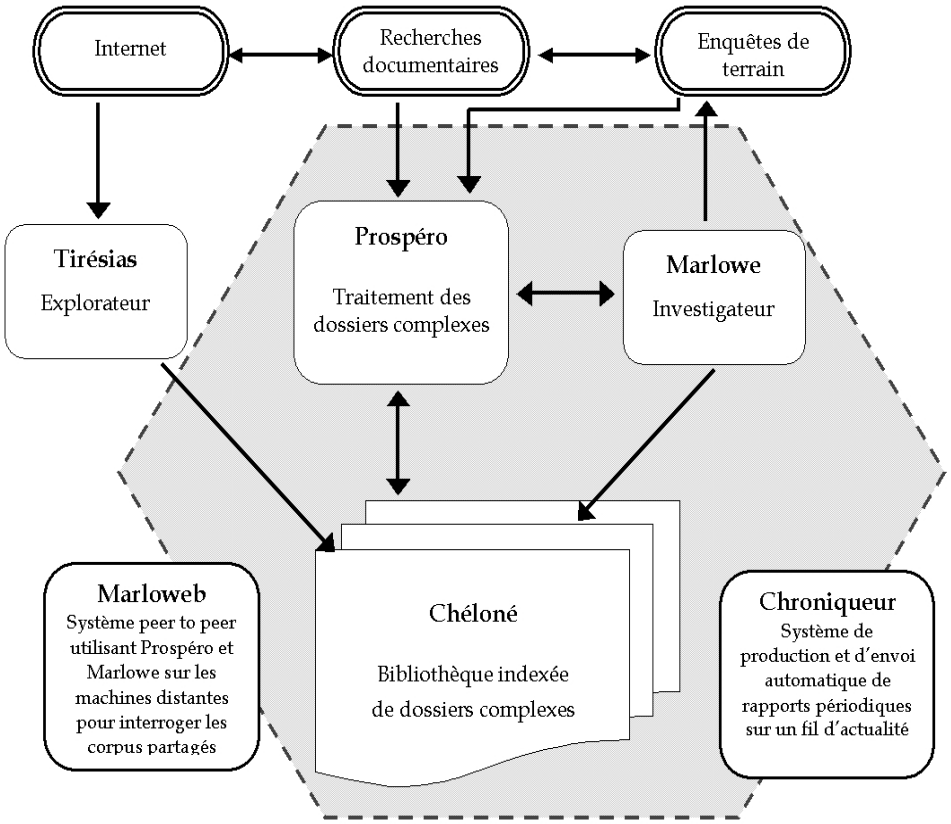
Rééquiper le laboratoire sociologique face à la prolifération des causes

Dans la figure idéale, la sociologie pragmatique suit sur la longue durée l'évolution de dossiers d'alertes, de controverses, d'affaires ou de conflits en examinant systématiquement les transformations des configurations d'acteurs et d'arguments. On raisonne à partir de la confrontation d'un double espace de variations : une série de dossiers (nucléaire, OGM, nanotechnologies, climat, etc.) ; un ensemble de formes de mobilisation et d'argumentation (affaires, débats publics, modes de protestation, procédures d'enquête et d'expertise, modes de « gouvernance », etc.). C'est ce qui se noue et se dénoue au croisement de ces deux plans qui fait l'objet de nos réflexions théoriques. Pourquoi la mobilisation est-elle si radicale sur tel dossier ? Comment la forme instituée du débat public peut-elle modifier des rapports de forces et de légitimités sur tel autre ? On postule que les confrontations d'acteurs et d'arguments sont créatrices d'espaces de possibles à partir desquels s'infléchissent et se réfléchissent les trajectoires des objets. Par exemple, une discussion sur l'énergie nucléaire produit une nouvelle hiérarchisation des jeux d'acteurs et d'arguments et, sans pour autant désigner un vainqueur définitif, permet de caractériser à la fois un rapport de forces et une configuration discursive dominante – dans une période donnée, on parle par exemple beaucoup plus des énergies que des risques sanitaires ou des conditions de sécurité dans les centrales. Lorsque l'on suit de tels processus, la vitesse de reprise, d'influence, de modification et de redistribution des points de vue et des coups que permet le Web, surtout depuis la généralisation du haut débit, change complètement les modalités de l'enquête. Du même coup, la tendance du chercheur est de suspendre la circulation et la prise en compte de multiples sources hétérogènes, pour se concentrer sur des sources plus stables, dotées de références ou de garanties de fiabilité, et généralement placées elles-mêmes en position de centre de tri ou de réduction des informations pertinentes sur un dossier. Ainsi, pour réduire les opérations visant à discerner l'émergence de nouvelles alertes dans le domaine de la santé environnementale, on a été conduit à se concentrer sur le site du Journal de l'Environnement, lequel a rapidement fait apparaître des biais liés aux préoccupations des acteurs porteurs du site. Ces biais ne sont apparus que parce que l'on pouvait croiser systématiquement des sources en transformant leurs informations en corpus comparables.

Pour parvenir à suivre et à analyser en profondeur les multiples corpus évolutifs qui intéressent notre sociologie, on s'est doté d'une architecture dans laquelle opèrent plusieurs entités logicielles. Sans redéployer ici les attendus de ce dispositif socio-informatique, on peut en rappeler le schéma général. Il s'agit de lier quatre fonctions décisives : la structuration des informations (ici des propriétés de grands corpus) ; l'analyse de ces informations (la mise en perspective de jeux d'acteurs et d'arguments à travers des textes et des énoncés, des classes d'objets et des formules) ; la mise à jour et le suivi de sources déterminées sur le Web (exploration de sites et de fils d'informations) ; et enfin l'archivage et la construction d'une mémoire dynamique qui sert de socle de référence pour le repérage et la caractérisation des nouvelles configurations et des processus émergents, ce qui permet de décliner le mot d'ordre selon lequel il n'y a pas

[6] Voir le modèle des « issue crawlers », qui renvoient les cartes de liens entre les sites ou les pages qui traitent des mêmes sujets sur la Toile, mais qui n'ont pas de corpus de référence ni de cadre d'analyse explicite. Voir le « Issue Crawler project » sur http://issuecrawler.net. N. Marres, "Tracing the trajectories of issues, and their democratic deficits, on the Web: The case of the Development Gateway and its doubles", Information Technology & People, 2004, Vol. 17, 2, p. 124 - 149

de veille sans mémoire[7]. L'implémentation de ces fonctions a pris corps dans quatre supports informatiques, formant la suite Prospéro, Marlowe, Tirésias et Chéloné.



Le logiciel Marlowe, qui nous intéresse plus particulièrement ici, est né au cours de l'été 1999. D'abord conçu comme un sous-programme du logiciel Prospéro, destiné à l'activation de fonctions spécialisées, il a pris au fil du temps de plus en plus d'autonomie, notamment depuis l'année 2003, au cours de laquelle a eu

[7] La question de la pérennité des sources est assez cruciale avec les évolutions rapides du Web. On lit des études assez pessimistes sur la capacité du Web à engendrer des archives durables et fiables. Voir W. Koehler, « Web page change and persistence - a four-year longitudinal study», Journal of the American Society for Information Science and Technology, v.53 n.2, p.162-171, January 15, 2002. La question des outils d'archivage était une des préoccupations majeures du dernier sommet de la société de l'information : "As increasing amounts of information find their ways into the Internet's archives, it is vital that we preserve their accessibility, renderability and interpretability. Digital documents often need to be interpreted by specific software packages to be rendered in understandable form. We will need to assure that the bits we preserve on digital media can also be read and understood not only by people but by computers programmed to help us manage this ocean of information. Steps are needed to assure that the information we accumulate today will be usable not merely decades but centuries and even millennia into the future. We need to preserve access to application software, operating systems and perhaps even hardware or simulators so as to retain the ability to make effective use of our digital archives". (Vint Cerf (Chairman, ICANN), "Governance of the Internet: the tasks ahead", Internet Governance Forum ;Athens, Greece, October 30, 2006).

lieu une performance publique mémorable[8]. A partir de 2004, Marlowe (acronyme MRLW) est intégré dans un réseau de recherches en sociologie et, tout en contribuant aux enquêtes, poursuit son « apprentissage ». Plusieurs textes ont déjà évoqué les premiers pas et les présupposés socio-logiques de ce personnage virtuel voué à affronter à la fois d'importantes turbulences dans le champ des sciences sociales contemporaines et des mutations considérables dans les modes de traitement de l'information[9]. En tant qu'enquêteur virtuel ou sociologue électronique Marlowe a besoin d'un réseau de chercheurs et de logiciels avec lesquels il peut développer pleinement une philosophie du dialogisme fondée sur la mise à l'épreuve constante de propositions qui n'ont pas besoin d'être consensuelles pour produire des connaissances[10]. Dans la panoplie des outils de Marlowe, la réalisation de chroniques occupe une place importante et c'est sur cet aspect que nous allons nous concentrer dans la suite du texte.

Les chroniques du logiciel Marlowe

Comme d'autres formes utilisées par Marlowe, telles que la citation, la définition, l'éphéméride, le proverbe, le résumé, l'anecdote, le récit d'expérience, la biographie ou la nécrologie, la chronique est un genre issu d'une longue histoire lettrée, bien décrite par l'histoire des technologies intellectuelles. Dès la fin du III^{e} millénaire, commence un travail d'archivage des oracles et des événements marquants auxquels ils renvoient. Parmi les érudits, l'idée s'impose alors que l'on peut connaître le passé et que l'avenir est prévisible. Vers 2200, sous le règne de Naram-Sîn d'Akkadé, des scribes composent une chronique royale, la première du genre selon Glassner[11]. La chronique a alors pour but de manipuler le passé afin d'expliquer le présent et de légitimer du même coup le pouvoir du souverain. Le genre de la chronique a ainsi une origine profonde, liée à la systématisation des usages de l'écrit et à l'apparition de classes de lettrés, qui vont progressivement s'attacher à rassembler les savoirs et organiser la mémoire collective, en créant les premières bibliothèques [12]. Evidemment, l'imprimerie jouera par la suite un rôle décisif dans les transformations du statut de l'écrit et du type d'autorité qui sera conféré à la faculté de croiser et de recouper les sources [13]. En Occident, après l'émancipation des lettrés de l'emprise de l'Eglise catholique, des chroniques de plus en plus spécialisées verront ainsi le jour. La naissance des revues savantes est le produit de ce long processus de spécialisation des savoirs et des formes d'érudition. En suivant l'évolution de la relation critique aux technologies de pouvoirs-savoirs, le genre des chroniques s'est ainsi déployé tout au long des siècles. La standardisation du genre semble parachevée avec la professionnalisation de la presse écrite et radiophonique, mais l'exercice du chroniqueur ne cesse de changer de format et d'amplitude, comme le montre l'avènement des blogs, qui lui assure une nouvelle apogée.

La fonction de chroniqueur vient donc se ranger à la suite d'une longue tradition, fortement marquée par le genre critique et polémique[14]. De nos jours, le ton d'une chronique est celui d'un commentaire enlevé jouant de multiples ressorts rhétoriques tout en gardant une prise relativement ferme sur un événement dont le chroniqueur cherche à mettre en variation la véritable signification. De ce point de vue la chronique

---

[8] http://prospero.dyndns.org:9673/prospero/acces_public/02_textes_sur_prospero/id_info_performance_marlowe/id_performance_marlowe

[9] Voir notamment F. Chateauraynaud, « Marlowe - Vers un générateur d'expériences de pensée sur des dossiers complexes », Bulletin de Méthodologie Sociologique, n° 79, juillet 2003.

[10] Sur la philosophie dialogique de la connaissance, voir M. Beller, Quantum Dialogue. The Making of a Revolution, Chicago, The University of Chicago Press, 1999.

[11] Etudiant la chronographie en Mésopotamie antique, Glassner a montré comment l'organisation de l'écriture cunéiforme a pris appui sur une organisation politique, laquelle, pour produire de l'ordre, doit se doter d'interprètes, les devins, capables de lire dans les objets et les signes ce qui a trait au comportement humain - les dieux s'exerçant à distribuer les indices sur toutes sortes de supports. Voir les Chroniques mésopotamiennes, présentées et traduites par J.-J. Glassner (Paris, Les Belles lettres, 1993).

[12] Voir C. Jacob, « Rassembler la mémoire. Réflexions sur l'histoire des bibliothèques », Diogène, 2001/4 - n° 196, p. 53 à 76.

[13] E. Eisenstein, La révolution de l'imprimé à l'aube de l'Europe moderne, Paris, La Découverte, 1991.

[14] Voir M. Angenot, Dialogues de sourds. Traité de rhétorique antilogique, Paris, Mille et une nuits, 2008.

entretient un rapport intime avec la veille : soumis à une contrainte d'actualité, le genre est associé à l'idée d'une sélection percutante de traits ou d'aspects qui permettent une mise en perspective d'un événement ou d'un discours et son insertion dans une série historique, fut-elle rappelée à des fins ironiques. La chronique sert d'outil de veille collective en ce qu'il rend saillants les rapports entre des séries passées, des événements ou des propos présents et des séries à venir dont elle vise à cerner les potentialités. De ce point de vue c'est un dispositif performatif qui pèse sur les représentations que se font ses lecteurs ou ses auditeurs des tendances, des continuités ou des ruptures. En dépit de quelques traits d'humour stylisés et faciles à programmer à partir de mises en variation langagières, le modèle expérimenté dans Marlowe est plus « plat » que les chroniques que l'on trouve dans les medias ou sur les blogs. Il ne s'agit pas à proprement parler d'un point de vue subjectif même si le système dispose d'une large ouverture des possibles par la mise en concurrence d'une pluralité de scripts. A la simulation de chroniques humaines, qui était une des options possibles, on a préféré la mise en œuvre systématique, ou plutôt la recombinaison, de routines déjà incorporées dans le couple Prospéro-Marlowe, de façon à conserver une unité de style : il ne s'agit pas de singer des chroniqueurs humains déterminés mais bien de développer un ensemble de variantes adaptées aux capacités propres de la machine. Si la dimension « exercice de style » est présente, c'est en mode mineur, du moins pour l'instant. Cela dit, la forme chronique est en soi un système de production sous contrainte qui rend possibles d'infinies variations. La réalisation d'une chronique hebdomadaire sur le fil santé-environnement dans le cadre d'une convention avec une agence comme l'AFSSET[15] fournit l'occasion de concrétiser cette fonction de chroniqueur.

Genèse et structure d'une chronique automatique très spécialisée

Dans le cadre de la poursuite d'un projet d'observatoire sociologique informatisé, il est apparu pertinent de se doter d'un outil de veille capable d'adresser, selon une fréquence hebdomadaire, une chronique sanitaire et environnementale, à un réseau de partenaires. Cette chronique hebdomadaire spécialisée a été conçue par adaptation d'une chronique quotidienne généraliste mise en place à la fin de l'année 2004 et en fonctionnement continu depuis. Ce dispositif expérimental n'a pas pour but de produire un outil de communication automatique diffusable tel quel à l'extérieur, mais de s'assurer que l'ensemble des protocoles d'indexation et de codage du contenu des séries textuelles est transposable à des informations en flux non supervisées par des interprètes humains. A un premier niveau, ce dispositif permet de prendre note de la progression des dossiers (objets d'alertes) dans l'actualité sanitaire et environnementale. Si le taux d'échec des procédures analytiques de Marlowe est relativement faible, la difficulté réside dans le passage d'un protocole d'acquisition de connaissances à une boucle d'apprentissage permettant au système de mieux sélectionner et structurer les informations qu'il mobilise[16]. Du point de vue de l'utilisateur, les chroniques permettent d'imaginer et de tester collectivement de nouveaux scripts transposables le logiciel Marlowe.

Comment fonctionne techniquement ce dispositif ? Le chroniqueur automatique repose sur la coopération de plusieurs instances : un module du logiciel Tirésias va sur la Toile chercher les informations de la semaine, tandis qu'un autre les transforme en corpus directement analysable sous Prospéro. Marlowe entre en lice en bout de course : disposant d'une pluralité de scripts et de chemins possibles (voir un état de l'organigramme en annexe), il extrait des informations qu'il juge pertinentes et organise leur « commentaire ». Une fois qu'il a terminé la rédaction de sa chronique, il l'adresse par courriel à un réseau de collaborateurs humains. Dans la foulée, il met à jour des fichiers qui lui servent à établir des palmarès et à repérer des évolutions, à construire pas à pas la structure de ce que l'on peut attendre dans l'actualité, à établir des listes de traits qu'il s'efforce de classer. Il crée également une base d'informations marquantes dont il pourra se servir ultérieurement sans avoir besoin de consulter de nouveau la base complète des dépêches.

---

[15] Cette convention de recherche passée entre l'Agence Française de Sécurité Sanitaire de l'Environnement et du Travail et l'EHESS portée par le GSPR en partenariat avec l'association Doxa a débuté en 2006 et a été renouvelé en 2007 jusqu'en 2010.

[16] On a joint en annexe de larges extraits d'une chronique adressée par Marlowe le 6 octobre 2008. Le « corpus » des chroniques disponibles pour évaluer le fonctionnement du dispositif est d'ores et déjà immense puisque nous disposons de plus de 1400 chroniques quotidiennes et d'une centaine de chroniques hebdomadaires sur le fil santé-environnement.

Soit un exemple de structure élémentaire de données générée par le chroniqueur hebdomadaire : la table des « objets d'alerte » de la semaine – ce genre de table est alimenté depuis fin 2006 (il s'agit ici d'une courte sélection depuis septembre 2008)

1/ 9/2008 - 7/ 9/2008 : ouragan nucléaire inondations CO2 pluies torrentielles cyclone gaz à effet de serre plomb glissements de terrain radon bruit couche d'ozone déforestation changement climatique radiothérapie nosocomiales déchets
8/ 9/2008 - 14/ 9/2008 : ouragan pesticides déchets CO2 changement climatique radiothérapie sécheresse inondations gaz à effet de serre OGM nanoparticules attentat déforestation dioxyde de carbone nucléaire terrorisme vache folle bruit poussière air intérieur pollution de l'air monoxyde de carbone composés organiques volatils champs électromagnétiques
15/ 9/2008 - 21/ 9/2008 : déchets couche d'ozone changement climatique bruit CO2 plomb saturnisme pesticides cyclone amiante incendies sécheresse gaz à effet de serre antennes-relais arsenic mercure incendie hormones tabac nuisances sonores changements climatiques perturbateurs endocriniens grippe aviaire marées noires OGM nitrates marée noire métaux lourds chlore
22/ 9/2008 - 26/ 9/2008 : CO2 déchets paludisme déforestation sida OGM mercure arsenic plomb pesticides gaz à effet de serre sécheresse bruit changement climatique greffes saturnisme canicule alcool tabac eaux de baignade pollution de l'air maladies cardio-vasculaires antennes-relais radioactivité poison nucléaire chlore
29/ 9/2008 - 5/10/2008 : déchets amiante dioxines pesticides CO2 pollution de l'eau bruit changement climatique espèces menacées gaz à effet de serre éthers de glycol nucléaire inondations cigarette benzène formaldéhyde sécheresse radon alcool pollution atmosphérique pollution de l'air SO2 NOx rayonnements ionisants métaux lourds radioactivité mercure plomb
6/10/2008 - 12/10/2008 : amiante déchets changement climatique CO2 nitrates gaz à effet de serre VIH nanoparticules tabac grippe aviaire sida attentats tabagisme pollution de l'air H5N1 fièvre jaune Ebola nucléaire terrorisme incendie cigarette espèces menacées poussière effet de serre pollution atmosphérique épizootie monoxyde de carbone NOx OGM pyralène benzène dioxyde de carbone pesticides
6/10/2008 - 12/10/2008 : amiante déchets changement climatique CO2 nitrates gaz à effet de serre VIH nanoparticules tabac grippe aviaire sida attentats tabagisme pollution de l'air H5N1 fièvre jaune Ebola nucléaire terrorisme incendie cigarette espèces menacées poussière effet de serre pollution atmosphérique épizootie monoxyde de carbone NOx OGM pyralène benzène dioxyde de carbone pesticides
13/10/2008 - 19/10/2008 : déchets pesticides CO2 cigarettes champs électromagnétiques tabac gaz à effet de serre plomb tsunami eau du robinet amiante UV poussière VIH antennes-relais nanoparticules rayons cosmiques pollution de l'air changement climatique hépatites hépatite C hépatite B Plomb OGM éthers de glycol arsenic toluène chlore

L'intérêt majeur de cette expérience est de fournir un banc d'essai pour les outils d'indexation, de codage et d'inférence incorporés dans les logiciels et utilisés par ailleurs sur toute une gamme de corpus spécialisés. On peut notamment envisager de lier les données à traiter et les analyses du chroniqueur à des algorithmes de classification et de tri[17].

Au premier stade de l'expérience, on a choisi d'appliquer le chroniqueur à un site d'informations génériques sur l'environnement, consultable en ligne. Les protocoles de Tiresias ont ainsi été adaptés pour construire le socle des articles depuis 2004. On y a ensuite ajouté une sélection hebdomadaire de dépêches d'agences qui, contrairement à la série précédente prise dans son exhaustivité, sont filtrées à l'aide d'un protocole vérifiant la présence d'une liste de thèmes et d'expressions liés à des problématiques sanitaires et environnementales. Un seul de présence et de déploiement (au moins plusieurs éléments de chaque classe d'indices), est utilisé afin d'éliminer les très nombreuses dépêches d'actualité politique et sociale trop générale mais aussi d'actualité sportive ou financière. La seule mention des mots « santé » et « environnement » ne suffit pas à réaliser un filtre sémantique fiable, et on procède donc par ajustements successifs des critères du filtrage : on parle en effet d' « environnement concurrentiel » ou de « santé des entreprises », d' « alertes » et de dangers ou même de « toxicité » dans toutes sortes d'activité qui n'ont rien à voir avec le fil santé-environnement qui nous intéresse

[17] Voir J. Velcin and J.-G. Ganascia "Topic Extraction with AGAPE", Proceedings of the International Conference on Advanced Data Mining and Applications (ADMA), Harbin, China, 2007.

Pour introduire dans la chronique hebdomadaire des variations permettant d'ouvrir différents possibles et d'éviter la monotonie d'un séquençage qui resterait identique de semaine en semaine, tout en assurant à chacun de ces possibles un bon degré de pertinence par rapport aux textes de la semaine, on a organisé les scripts en suivant un « processus aléatoire éclairé », les scripts étant tirés au sort mais la plupart d'entre eux étant assortis de conditions relativement fortes. Concrètement, à chaque niveau où plusieurs scripts sont compossibles, le système en prend un au hasard et réalise des tests pour savoir si les caractéristiques des textes de la semaine lui permettent de l'activer eu égard aux contraintes assignées à ce script. Par exemple, Marlowe porte une attention particulière à la présence éventuelle de thématiques rares mais extrêmement pertinentes comme celle des « faibles doses » ou encore celle des « perturbateurs endocriniens » par exemple. Il croise la recherche d'éléments déjà connus avec des éléments nouveaux ou émergents, et c' nouveaux ou émergents, et c'est dans la rencontre, à chaque fois différente, des deux plans que se fait l'intérêt de la chronique. Lorsqu'une classe d'objets ou une relation est détectée, certaines rubriques sont privilégiées de façon à structurer le fil de la chronique. Mais cette structuration, comme on le voit dans l'exemple fourni en annexe, laisse une large place à l'imprévu et à ce qui n'a pas encore été indexé ou répertorié dans les « ontologies » ou catégories sémantiques du système. Autrement dit, comme dans Prospéro, il s'agit de faire parler doublement les textes : à travers des filtres sémantiques ayant nécessairement incorporé un point de vue et à travers des propriétés émergentes saisies via des grappes, des liens ou des rapprochements inédits. Si aucune des conditions préalables à la structuration sémantique de la chronique n'est réunie, le système tire au sort un autre script du même niveau, vérifie les contraintes, et le cas échéant se replie sur des scripts élémentaires. Diverses possibilités sont prévues à l'intérieur de chaque script, y compris, dans certains cas, l'absence totale des propriétés sur lesquelles il est centré. En effet, il faut s'assurer qu'à chaque niveau il y ait toujours au moins un script possible puisque l'enchaînement des tests dépend du succès de l'étape antérieure au moins pour les deux premiers niveaux.

Une chronique de Marlowe se présente ainsi comme l'enchaînement de requêtes qui lui permettent d'extraire, à partir d'un corpus de dépêches lues quotidiennement sur le Web, des traits suffisamment marquants pour donner lieu à des commentaires, lesquels sont en quelque sorte automatiquement contextualisés. Ces derniers peuvent varier depuis le commentaire d'humeur - « les humains se font encore la guerre ! » - jusqu'au micro-rapport dédié spécifiquement à des programmes de recherche, comme tout ce qui concerne les alertes et les controverses ou les formes de mobilisation autour des enjeux sanitaires, environnementaux ou technologiques – «aujourd'hui, le dossier des OGM a encore rebondi !». Du point de vue procédural, la réalisation des chroniques repose sur l'enchaînement de cinq opérations distinctes : chaque jour, en plusieurs étapes, un module spécialisé de Tirésias, un « crawler », va lire sur la toile des textes librement accessibles ; en fin de journée, un autre module de Tirésias, dénommé le « veilleur », fabrique un corpus à partir des documents retenus – certains fils sont écartés, comme par exemple les résultats sportifs ou les événements culturels, ce qui ne va pas de soi mais procède d'un choix d'orientation du chroniqueur de Marlowe – le sport ou la culture ne surgissant que lorsqu'ils sont liés à des enjeux politiques, des mobilisations ou des affaires, ou encore des innovations technologiques ; un troisième module, appelé le « lanceur », qui sert habituellement à réitérer des questionnements en boucle sur plusieurs corpus, indexe les documents retenus et construit une base de connaissances sous Prospéro ; après quoi, à l'aide d'une centaine de scripts différents, Marlowe analyse les propriétés de ce corpus et en tire un certain nombre d'observations ; enfin, un cinquième module assure la transmission de la chronique au carnet d'adresse du logiciel.

Le cahier des charges de ce chroniqueur s'est fixé graduellement et, depuis le début de l'année 2005, Marlowe produit quotidiennement un texte qu'il adresse à un réseau d'interlocuteurs proches qui contribuent en retour à enrichir ses « scripts ». Peu coûteuse en terme de développement stricto sensu, la réalisation de la chronique quotidienne a permis d'avancer sur plusieurs lignes de front :

- En premier lieu, il s'agissait d'assembler des modules et des fonctionnalités déjà réalisés pour créer un dispositif autonome. L'autonomie est entendue ici au sens où, pour effectuer toutes les opérations qui mènent de la capture d'informations sur le Web à l'envoi d'un compte rendu rédigé à l'intention d'un réseau de lecteurs, ce dispositif ne suppose aucune intervention humaine. D'un point de vue formel, le procédé est

assez simple puisqu'il s'agit de projeter des éléments extraits de fils d'actualité, triés à l'aide d'indices élémentaires, dans des figures de style d'inspiration plus littéraire, en réalisant la jonction entre deux techniques d'écriture. La réalisation de ce dispositif, d'abord conçu sur un mode généraliste, a permis de concevoir des variantes tournées vers des problématiques ou des sources plus spécifiques. Il s'est avéré utile, par exemple, de développer un chroniqueur hebdomadaire spécialisé en matière de relations entre la santé et l'environnement.

- En deuxième lieu, ce processus rend clairement visibles, et partant critiquables, les procédés syntaxiques et sémantiques sur lesquels reposent les routines du système. Le fait de laisser une machine produire une interprétation de séries de documents au format relativement stable mais au contenu ouvert sur toutes sortes d'événements et d'affaires publiques fournit en effet un contre-test précieux face au travail plus classique d'ancrage de jeux de descripteurs et de catégories d'analyse sur des dossiers pris un à un et soumis à l'examen continu d'un expert humain. Les points de décrochage ou de dérivation sont plus clairement lisibles et permettent de poser la question du domaine d'extension d'un langage de description, dont l'analyse de multiples dossiers montre qu'il est soit trop général, soit trop spécialisé, et qu'une des qualités majeures du raisonnement humain, même sous forme de routines, est de négocier constamment les changements d'échelles ou de niveaux logiques.

- Enfin, le fait de disposer d'un agent capable de suivre continûment un flux d'informations permet de donner corps à l'idée de délégation socio informatique et d'assurer une mémoire transversale, fort utile pour recontextualiser des alertes ou des controverses et toutes sortes d'événements dont le caractère marquant n'est aperçu qu'après coup. Par exemple, en suivant quotidiennement l'actualité politique, en relevant systématiquement les thèmes liés au domaine des risques ou en repérant les affaires dont on parle le plus, le système peut générer une base d'événements marquants utilisables comme points d'appui pour contextualiser ou interpréter des transformations observées dans des séries particulières, coupées artificiellement du fonds avec lequel elles communiquent. On peut penser à ce que les politistes appellent, sans trop s'interroger sur la dimension fonctionnaliste de cette appellation, l' « agenda politique » : si des mobilisations ou des interventions ont lieu sur un dossier comme le nucléaire, il peut être utile de savoir si c'est, ou non, en période de campagne électorale ou lors de l'examen d'un texte de loi concernant l'énergie. Tout peut servir de contexte à tout et c'est bien la difficulté de la modélisation de l'interprétation de séries textuelles sur laquelle ont buté de multiples approches. Sans chercher à tout résoudre, on s'est donné pour objectif d'accroître la modularité du dispositif en développant pleinement l'idée d'une pluralité de technologies littéraires.

Forger des alliances interdisciplinaires pour explorer les différentes formes d'apprentissage du système

Le dispositif présenté ici peut-il relever de l'apprentissage ? La notion d'« apprentissage » a déjà une lourde histoire dans le champ de l'IA : dans la version radicale, apprendre à des machines à apprendre, ce n'est pas seulement les doter d'outils de « data mining », « d'indexation », ou encore de « clustering », c'est littéralement créer des calculateurs autonomes capables de revoir leur propre programme ; dans la version plus académique, c'est avant tout l'explicitation des boucles logiques et sémantiques que peut prendre en charge un programme[18]. Le nombre d'expériences est immense mais la controverse est toujours ouverte : la part d'ajout humain dans l'augmentation des capacités cognitives des programmes informatiques est toujours plus grande, de sorte que la question de leur faculté d'apprentissage est de plus en plus difficile à évaluer sans métaphore, comme dans le cas des « consciences artificielles » réputées « émergentes » selon Rey Kurzweil ou « constructibles » selon Alain Cardon[19]. L'idée que l'on est entouré d' « agents intelligents » de plus en plus autonomes circule sans que l'on puisse sérieusement mettre à l'épreuve les protocoles[20]. Pour ce

[18] Y. Kodratoff, Leçons d'Apprentissage Symbolique Automatique, Toulouse, Cepadues, 1988.
[19] R. Kurzweil, The Age of Spiritual Machine - When Computers exceed Human Intelligence, Penguin Books, New York, 1999 ; A. Cardon, Conscience artificielle et systèmes adaptatifs, Paris, Eyrolles, 1999.
[20] Voir les fils d'informations et les chroniques du site automatesintelligents.com

qui nous concerne, à savoir l'intelligence sociologique des dossiers complexes, on est conduit à rester beaucoup plus modestes : l'intelligence est d'abord dans les discours et les textes étudiés. Et la révision des connaissances et la dynamique inférentielle se produisent dans les textes eux-mêmes, et la première tâche consiste à apprendre à suivre les raisonnements des acteurs-auteurs. Mais à partir des cadres conceptuels et des premières procédures mis en place, il est possible d'interconnecter la suite logicielle Prospéro-marlowe avec des approches de clustering automatique et des protocoles d'évaluation des informations et de leurs changements graduels de statut, comme lorsque des micro-configurations basculent du mode mineur –on en cause dans les coins ou toujours à propos d'un domaine précis – au mode majeur – émergence d'un nouveau prototype ou d'un paradigme, comme ce fut le cas avec le principe de précaution, et aujourd'hui avec le modèle d'expertise du GIEC sur le climat transformé en modèle (cf. le Grenelle de l'environnement ..).

Un des antidotes à la dispersion, la redondance et la prolifération des noeuds et des liens qui caractérisent la mise en réseau non seulement des informations mais aussi des analyses de ces informations, consiste à se doter d'instruments alternatifs permettant de sérier les problèmes en prenant à la lettre la signification du verbe « sérier » : construire des séries raisonnées, c'est-à-dire dont la raison est explicitée et dont on peut fournir le principe de mise en équivalence ou d'homogénéité, ou encore de congruence. En concevant une collection de dossiers sous la forme de séries de corpus distribués entre de multiples machines et accessibles à un réseau d'enquêteurs capables non seulement d'enrichir les corpus, mais aussi de développer des outils d'analyse (descripteurs, catégories, formules, scripts et protocoles), on tente de retourner le rapport de force en remettant des contraintes cognitives fortes sur le traitement des séries et leurs comparaisons. Il s'agit ainsi d'assurer une explicabilité et une lisibilité des ressorts interprétatifs en vertu desquels des modèles, des concepts, des énoncés ou des théories réussissent mieux que d'autres à rendre compte d'un phénomène ou d'un processus. Ce déplacement ne rompt pas les liens avec l'exigence d'ouverture impliquée dans l'idée d'une connaissance ouverte et d'un Web coopératif par nature, mais oblige à sortir quelques temps de la boucle interprétative pour mettre à distance certaines classes d'événements ou d'énoncés sur ces événements, et en travailler la description, la compréhension et l'analyse avec d'autres outils que ceux que partagent les protagonistes eux-mêmes.

Annexe : Extrait d'un chronique santé-environnement exemplaire de Marlowe, le 6 octobre 2008

Le texte qui suit est le contenu assez typique d'une chronique hebdomadaire adressée automatiquement par Marlowe à une série de correspondants dans le but d'aiguillonner la veille sociologique collective sur le fil santé-environnement et de permettre la réinterrogation des corpus qui sous-tendent le dispositif et qui sont régulièrement mis à jour. Marlowe simule une sorte d'interrogatoire en s'adressant à lui-même des requêtes. Pour faire court et pour limiter le matériel au strict nécessaire, on a retiré des références littéraires utilisées notamment dans la phase d'amorce de la chronique. Pour « gratuites » qui puissent paraître ces références, elles ont pour fonction d'ancrer le dispositif dans une solide culture en sciences humaines et sociales …

[…]

Marlowe: lance le chroniqueur santé environnement

Marlowe : On a tous nos petites manies. En ce moment je scrute la moindre apparition de la problématique des "faibles doses". Cette semaine, je n'ai pas trouvé grand chose sauf dans le texte intitulé " Cancers et environnement: forts soupçons sur les pesticides " (JDLE 3/10/2008) :

Les experts de l ' Inserm ont insisté sur l ' importance de poursuivre les efforts de recherche pour améliorer les connaissances sur l ' effet à long terme des faibles doses de polluants, en jeu notamment pour les dioxines et les rayonnements ionisants, ainsi que l ' impact des " mélanges " de plusieurs facteurs ainsi que la susceptibilité individuelle face au risque.
Auteur :JDLE Date :03/10/2008 Titre :Cancers et environnement: forts soupçons sur les pesticides

Certes, ce n'est pas ça qui va faire faire un grand bond en avant à nos recherches sur la question des faibles doses mais je tenais à signaler que je suis vigilant !

Marlowe: approfondis

Marlowe : Si on regarde les objets d'alertes de la semaine, on obtient une liste ordonnée qui a la forme suivante :
déchets, amiante, dioxines, pesticides, CO2, pollution de l'eau, bruit, changement climatique, espèces menacées, gaz à effet de serre, éthers de glycol, nucléaire, inondations, cigarette, benzène, formaldéhyde, sécheresse, radon, alcool, pollution atmosphérique, pollution de l'air, SO2, NOx, rayonnements ionisants, métaux lourds, radioactivité, mercure, plomb …

J'ai entr'aperçu soigneusement cette liste et après avoir comparé consciencieusement, pour chaque objet d'alerte, les thèmes qui lui sont associés cette semaine à ceux qui forment son réseau typique dans l'ensemble du « corpus socle », j'ai retenu « cigarette ». En effet, la configuration engagée s'écarte du noyau habituel : décès, charbon, bois, chauffage, conjoint, usage, responsabilité, cause, poumon, cancers, mortalité, chroniques, respiratoires, maladies, Chine, renoncement, personnes, nombre, Etats-Unis, Harvard, santé publique, Ecole, Ezzati, Majid, Lin, Hsien-Ho, équipe, chercheurs …

Parmi ces éléments, 27 sont assez inhabituels : décès, charbon, bois, chauffage, conjoint, usage, responsabilité, cause, poumon, cancers, mortalité, chroniques, respiratoires, Chine, renoncement, personnes, nombre, Etats-Unis, Harvard, santé publique, Ecole, Ezzati, Majid, Lin, Hsien-Ho, équipe, chercheurs

Tant que j'y suis, je fournis la liste ordonnée des thèmes les plus fortement associés de manière générale (dans le socle) à un objet comme « cigarette » : alcool, risques, euros, dangers, tabac, industrie, réchauffement climatique, désinformation, campagne, ExxonMobil, pétrolier, stress, drogue, violence, BIT, sida, mégots, plastiques, matières, voitures, carcasses, fumée, perturbateurs endocriniens, monoxyde de carbone, bactéries, amiante, santé, effets, maladies, composés organiques volatils, ventilation, bâtiments, âge, activités, chant, facteurs aggravants, scientists, Union, association, bénévoles, Australie, Nettoyons, la journée, année, succès, ambition, international, bâtiment, syndrome, substance …

Un peu de verbatim rendra tout ceci plus cristallin :

Les chercheurs, dont l ' équipe était conduite par Hsien-Ho Lin et Majid Ezzati de l ' Ecole de santé publique d ' Harvard (Etats-Unis), ont estimé le nombre de personnes qui devraient mourir d ' ici 2033 en Chine de maladies respiratoires chroniques (2e cause de mortalité) ou de cancers du poumon (6e cause), et la responsabilité dans ces décès de l ' usage conjoint de la cigarette et du charbon ou du bois de chauffage. Auteur :AFP Date :04/10/2008 Titre :Chine: tabac et chauffage domestique feront 32 millions de morts d'ici 2033

Ils ont pu établir ainsi que le renoncement à la cigarette et en même temps au chauffage au charbon ou au bois devrait éviter 32,3 millions de décès.
Auteur :AFP Date :04/10/2008 Titre :Chine: tabac et chauffage domestique feront 32 millions de morts d'ici 2033

[…]

L'aventure de cette chronique environnementale et sanitaire (et vice versa) ne faisant que débuter, je pioche un peu dans des répertoires d'objets convenus. Ainsi, je note la présence de thèmes comme PNSE, expertise, exposition professionnelle, santé publique, développement durable, principe de précaution ... Voici de quoi justifier (éventuellement) cette sélection :

PNSE : *L ' Afsset utilisera également ce rapport dans le cadre de sa contribution au plan national Santé-environnement (PNSE 2), au plan Cancer 2 et au plan Santé au travail 2*. Auteur :JDLE Date :03/10/2008 Titre :Cancers et environnement: forts soupçons sur les pesticides

expertise : *(1) Cancer et environnement - expertise collective - Inserm, Afsset (octobre 2008)* Auteur :JDLE Date :03/10/2008 Titre :Cancers et environnement: forts soupçons sur les pesticides

exposition professionnelle : *L ' exposition professionnelle aux éthers de glycol et aux autres facteurs de risque a été évaluée à partir de questionnaires, d ' entretiens individuels et d ' une expertise par des spécialistes de l ' hygiène au travail*. Auteur :JDLE Date :30/09/2008 Titre :Ethers de glycol: facteur de risque d'infertilité pour l'homme

[…]

Marlowe : Depuis Duval R., Temps et vigilance, Paris, Vrin, 1990., on sait que les formes d'ouverture d'avenir se déclinent de plusieurs manières (le projet, le délai, l'attente, l'anticipation, l'agenda - j'abrège pour ne pas saturer la fenêtre -). Quod erat demonstrandum...

2010 : *La taxe poids lourd, toujours en discussion selon le secrétaire d'Etat au transport Dominique Bussereau, pourrait entrer en vigueur en 2011, et en 2010 en Alsace*. Auteur :JDLE Date :29/09/2008 Titre :Projet de loi de finances 2009: "un verdissement de la fiscalité"

D'après le projet de loi, la réduction de la taxe intérieure sur les produits pétroliers (TIPP), de 0,22 euro par litre actuellement pour le biodiesel, passera à 0,135 ¤ en 2009, à 0,10 en 2010, 0,06 en 2011 et sera abandonnée en 2012. Auteur :JDLE Date :02/10/2008 Titre :Biocarburants : suppression des aides fiscales d'ici 2012

Pour l ' éthanol, la réduction de TIPP (0,27 ¤ par litre actuellement) sera ramenée à 0,17 ¤ dès l'an prochain, puis 0,15 en 2010, 0,11 en 2011 et enfin zéro en 2012. Auteur :JDLE Date :02/10/2008 Titre :Biocarburants : suppression des aides fiscales d'ici 2012

2015 :
Le projet de loi de finances 2009 crée une taxe incinération et double, d ' ici 2015, la taxe décharge. Auteur :JDLE Date :01/10/2008 Titre :Evolution de la réglementation sur les déchets: des réactions mitigées

Boom du numérique aidant, la consommation européenne de ces équipements devrait passer à 14 TWh/an en 2015. Auteur :JDLE Date :01/10/2008 Titre :Eco-conception: les Etats membres approuvent la Commission

Réduction de 25 kg de la production de déchets en 5 ans ; recyclage de 35% des déchets ménagers et assimilés en 2012, puis 45% en 2015 (contre 24% en 2004); baisse de 15% des quantités incinérées ou mises en décharge en 2012. Auteur :JDLE Date :01/10/2008 Titre :Evolution de la réglementation sur les déchets: des réactions mitigées

[…]

Un cadre formel pour la veille numérique sur la presse en ligne

Mathilde Forestier*, Julien Velcin *,**
Jean-Gabriel Ganascia ***

*laboratoire ERIC, Université Lumière Lyon 2,
Faculté de Sciences Économiques et de Gestion
Département d'Informatique et de Statistique
5 avenue Pierre Mendès-France
69676 BRON Cedex
mathilde.forestier@eric.univ-lyon2.fr
**julien.velcin@eric.univ-lyon2.fr
http://eric.univ-lyon2.fr/~jvelcin/
***LIP 6,
104 avenue du Président Kennedy
75016 Paris
Jean-Gabriel@Ganascia.name

Résumé. La presse représente une source de données très volumineuse, dotée d'une contextualisation temporelle importante que ce soit au niveau de l'événement ou plus largement des thématiques traitant des événements. Cet article amorce une réfléxion sur le suivi temporel des données textuelles. Nous présentons ici un cadre formel (à un niveau d'abstraction assez élevé) permettant le suivi de thématiques dans le temps tout en restant sensible à l'inattendu qui se cache dans les données. Ainsi, ce cadre formel a pour but de détecter les évolutions des thématiques regroupant les articles de presse pour fournir, à terme, un résumé temporel de l'actualité ainsi qu'un système d'alerte performant destiné à prévenir l'analyste des phénomènes de l'actualité.

1 Introduction

Les articles de presse sont une source de données intarissable. Mais ces données volumineuses s'accompagnent de nouvelles problématiques telles que la recherche d'information ou encore la veille numérique. Bien que la veille soit un terme générique associé à un bon nombre de domaine (technologique, économique, etc.), nous entendons par *veille numérique* tout ce qui a trait aux technologies de l'information et communication (TIC) (Dumas et Bertacchini (2001)). De fait, nous travaillons exclusivement sur des données numériques (exclusivement textuelles) qui proviennent d'internet.

Des milliers d'articles paraissent chaque jour sur internet, ces articles peuvent être regroupés dans des thématiques (groupe d'articles sémantiquement cohérent) dans le but d'aider l'utilisateur à accéder aux articles qui l'intéressent mais aussi de synthétiser les données à analyser. Ainsi, que ce soit par un sytème a priori ou a posteriori, les thématiques ainsi formées sont donc (tout comme les articles) inscrites temporellement : nous pouvons parler aujourd'hui d'un sujet qui ne sera plus d'actualité demain. Les événements survenus en Birmanie sont un exemple probant. Nous en avons beaucoup parlé en septembre 2007 lors de la révolte des moines, puis la presse n'en a plus fait echo jusqu'en mai 2008 où un cyclone a devasté une grande partie du pays rapportant à la une des journaux les problèmes liés à la dictature.

Ainsi, nous proposons un cadre formel permettant de suivre les thématiques dans le temps. Il se base sur différentes évolutions possibles des thématiques, que nous préciserons dans cet article afin de concevoir un résumé des informations temporelles. De plus, ce cadre intègre la notion de veille et plus particulièrement celle d'inattendu (intrinsèquement liée à la veille) : il ne suffit pas de rechercher ce que l'on est de sûr de trouver (notamment par un grand nombre de mots communs, etc.) mais aussi ce que l'on n'espère pas (ce qui implique de nouvelles méthodes de fouilles de textes). Un système de détection d'événements et d'alerte a été élaboré pour prevenir l'analyste des changements dans l'actualité.

La section 2 présente un court état de l'art relatif à l'extraction et au suivi de thématiques à partir de la presse. Après avoir évoqué quelques courants en matière de suivi de thématiques, nous proposerons un cadre formel que nous avons élaboré, ainsi qu'une expérimentation réalisée à partir de données réelles.

2 État-de-l'art

L'état-de-l'art que nous présentons ici fait référence a deux aspects essentiels de notre travail à savoir tout ce qui a trait à la veille, à l'extraction et au suivi temporel des thématiques.

2.1 Veille

La veille peut avoir plusieurs définitons, selon le domaine auquel elle est attachée. Que l'on parle de veille technologique, de veille stratégique, de veille numérique, la définition peut différer mais elle s'apparente toujours à une "analyse de l'environnement suivies de la diffusion bien ciblée des informations selectionnées et traitées" (Jakobiak (1991)). De plus, la veille a pour but de repérer l'inattendu, ce qui ne se voit pas, comme il est expliqué dans Jacquenet et al. (2004) " la veille introduit une difficulté inhabitulle par rapport aux domaines d'application classiques des techniques de fouille de textes, puisqu'au lieu de rechercher de la connaissance fréquente cachée dans les données, il faut rechercher de la connaissance inattendue". Ainsi, est inattendu ce qui n'est encore jamais apparu (dans notre problème, une nouvelle thématique par exemple), ou bien ce qui disparaît brutalement etc. Dans Jacquenet et al. (2004), le logiciel Unexpected-Miner compare les nouveaux documents aux anciens enregistrés dans la base. Si le nouveau document ne parle d'aucune des thématiques de la base de donnée, alors il exprimera une nouveauté donc de l'inattendu. Toutefois, comme il est stipulé dans Jacquenet et Largeron (2005), les logiciels d'extraction existant sur le marché répondent assez mal aux besoins du veilleur pour deux raisons principales : soit ces logiciels requièrent un paramétrage trop lourd soit une connaissance précise du domaine par le veilleur.

2.2 Extraction et suivi de thématiques

TDT : Topic Detection and Tracking (Allan et al. (1998), Allan (2002), Binsztok et Gallinari (2002), Schultz et Liberman (1999))

Le TDT est un programme de recherche collaboratif créé en 1997 à la demande de la DARPA [1]. Ce programme réunit quatre équipes et a pour but trouver un système robuste et performant capable d'extraire automatiquement les thématiques dans un flux de données. Ils ont défini trois notions : l'événement (un fait à un moment et un endroit précis), l'histoire (toutes les nouvelles qui traitent d'un même événement) et la thématique (toutes les histoires qui discutent d'événements similaires). Ces trois définitions permettent la création de systèmes développés autour de 3 tâches principales. La première tâche est la segmentation d'un flux continu en dépêches indépendantes (traitant d'un évènement). Ensuite, vient la tâche de détection qui, comme son nom l'indique, détecte dans le flux un nouvel évènement qui n'a encore jamais été traité. Enfin, la tâche de suivi permet de classer les nouvelles relatant un événement déjà connu.

Text-Mining Temporel Le Text-Mining temporel (Mei et Zhai (2005)) associe à une thématique un ensemble de *Theme span* inscrits temporellement. Un *Theme span* est défini comme une distribution probabiliste de mot caractérisant une thématique ou une sous thématique sémantiquement cohérente. Ainsi chaque période sera définie par un nombre fixé de Theme span reliés entre eux par des flèches décrivant leurs relations d'évolutions. Enfin, la détection de nouvelles tendances (Kontostathis et al. (2003)) recherche à partir de mots clés l'émergence de nouveau courant scientifique (XML par exemple).

Ces deux approches sont sans conteste en relation étroite avec le travail présenté dans cet article. Toutefois, la notion des différentes évolutions est peu ou prou prise en compte, ce qui est un des aspects essentiels de notre travail. De même, la notion de veille correspond bien à une partie de notre problématique mais les techniques utilisés ne sont pas adéquates. En effet, l'information inattendue que nous voulons détecter ne doit pas être induite par les connaissances de l'utilisateur sur un domaine précis mais au contraire survenir à partir des données elles-mêmes.

3 Cadre Formel / Architecture

Nous considérons les articles de presse comme des sacs de mots pondérés par la mesure TF-IDF, qui est une représentation classique utilisée dans le domaine de la fouille de texte (Steinbach et al. (2000)).

[1] Defense Advanced Research Projects Agency, http ://darpa.mil

3.1 Evolution de thématiques

Notre cadre est basé sur la théorie des ensembles. Nous nous restreignons ici à deux périodes mais il est évident que le système permettra d'en gérer un plus grand nombre. Étant donnés (cf. figure 1) :

- **P** : la période de récupération des articles de presse. Cette période peut être exprimée à différents niveaux de granularité : une heure, un jour, une semaine, etc.
- **N** : l'ensemble de tous les articles possibles.
  - l'ensemble de tous les articles parus à la période p : N_p
  - l'ensemble de tous les articles parus à la période p+1 : N_{p+1}
  - On a bien entendu : $N_p \subset N$, $N_{p+1} \subset N$
- **T** : l'ensemble de toutes les thématiques possibles.
  - L'ensemble de toutes les thématiques extraites à partir de N_p : T_p y compris l'ensemble vide
  - L'ensemble de toutes les thématiques extraites à partir de N_{p+1} : T_{p+1} y compris l'ensemble vide
  - De même on a : $T_p \subset T$, $T_{p+1} \subset T$. Notons qu'il est possible de retrouver deux thématiques identiques à deux périodes différentes.

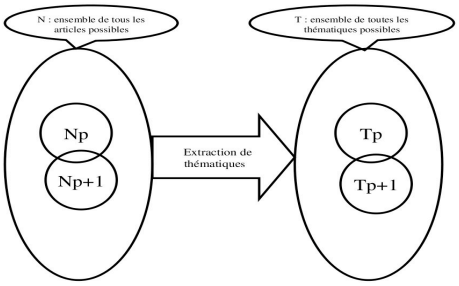


FIG. 1 – *Représentation des ensembles*

A partir de ce cadre, nous pouvons définir une relation R telle que $\forall\ x\ \in\ T_p,\ \forall\ y\ \in\ T_{p+1}$:

- xRy si sim(x,y) > α tel que sim(x,y) calcule le degré de similarité entre x et y compris entre 0 et 1 et α est un réel utilisé comme seuil.
  - Si $\forall\ x\ \in\ T_p,\ \nexists\ xRy\ \Longrightarrow \emptyset Ry$
  - Si $\forall\ y\ \in\ T_{p+1},\ \nexists; xRy\ \Longrightarrow xR\emptyset$
- R est la réunion de toutes les sous relations : $R = RT \cup RF \cup RS \cup RA \cup RD$ que nous définissons ci-dessous grâce à logique des prédicats du premier ordre et qui correspond aux différentes évolutions possibles (cf. figure2).

Transformation (Notée RT) : une thématique à la période p existe encore à la période p+1. Nous définissons le prédicat $transformation_p(x, y)$ comme vrai si xRT_py avec : $RT_p = \{(x, y) \in T_p \times T_{p+1}\}$

- xRy
- $\forall x, x' \in T_p, [(xRTy) \wedge (x'RTy)] \Rightarrow (x = x')$
- $\forall y, y' \in T_{p+1}, [(xRTy) \wedge (xRTy')] \Rightarrow (y = y')$

Fusion (Notée RF) : au moins deux thématiques à la période p vont fusionner en une thématique à la période suivante. Nous définissons le prédicat $fusion_p(x, y)$ comme vrai si xRF_py avec : $RF_p = \{(x, y) \in T_p \times T_{p+1}\}$

- xRy
- $\forall x \in T_p, \forall y, y' \in T_{p+1}, [(xREy) \wedge (xREy')] \Rightarrow (y = y')$
- $\exists x' \in T_p,\ avec\ x \neq x'$ tel que $x'Ry$

Eclatement (Notée RS) : une thématique à la période p va éclater en au moins deux thématique à la période suivante. Nous définissons le prédicat $split_p(x, y)$ comme vrai si xRS_py avec :

$RS_p = \{(x, y) \in T_p \times T_{p+1}\}$

- xRy
- $\forall x, x' \in T_p, \forall y \in T_{p+1}, [(xREy) \wedge (x'REy)] \Rightarrow (x = x')$
- $\exists y' \in T_{p+1},\ avec\ y \neq y'$ tel que xRy'

Apparition (Notée RA) : une nouvelle thématique va apparaître à la période p+1. Nous définissons le prédicat $appearance_p(\emptyset, y)$ comme vrai si $\emptyset RA_p y$ avec :
$RA_p = \{(\emptyset, y) \in T_p \times T_{p+1}\}$ et $\emptyset Ry$

Disparition (Notée RD) : une thématique à la période p n'existera plus à la période suivante. Nous définissons le prédicat $disappearance_p(y, \emptyset)$ comme vrai si $xRD_p\emptyset$ avec :
$RD_p = \{(x, \emptyset) \in T_p \times T_{p+1}\}$ et $xR\emptyset$

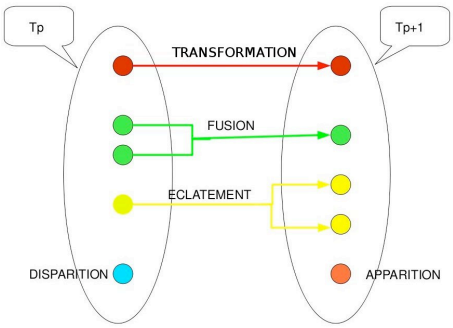


FIG. 2 – *Représentation des évolutions*

3.2 Alertes

La mise en place de ce cadre formel va permettre d'instaurer un certain nombre d'alertes qui vont prévenir l'analyste des événements de l'actualité. Pour cela, chaque thématique va être dotée d'un poids correspondant au nombre d'articles couvert par la thématique par rapport aux nombres d'articles total parus durant la même période. Ainsi, nous pouvons définir deux types de thématiques : les thématiques de faible poids ($weak\_topic$) où poids(x) < λ et les thématiques avec un poids assez élevé ($important\_topic$) où poids(x) > σ où λ et σ sont deux réels compris entre 0 et 1 avec $\sigma > \lambda$.

Transformation Nous pouvons définir deux alertes majeures pour l'opérateur de transformation : l'alerte de perte d'intérêt et celle du gain d'intérêt. L'alerte de gain d'intérêt va mettre en exergue une thématique qui va prendre de l'importance entre deux périodes. Cette alerte va être définie comme suit :

SI $transformation_p(x, y)$ ET $weak\_Topic(x, T_p)$ ET $important\_Topic(y, T_{p+1})$
ALORS ALERT(Gain_Interest,x)

L'alerte de perte d'intérêt va, à l'inverse, démontrer une perte de poids entre les deux périodes analysées.

Fusion Nous pouvons catégoriser deux types de fusion qui vont nous permettre de créer différentes alertes. La fusion par absorption va alerter l'analyste sur un fait de l'actualité qui va regrouper une thématique avec un poids élevé et une autre de poids faible. Ainsi la thématique importante va absorber une autre plus faible. Ce phénomène va être détecté comme suit :

SI $fusion_p(x, y)$ ET $fusion_p(x', y)$ ET $important\_Topic(T_p, x)$ ET $weak\_Topic(T_p, x')$ ET $important\_Topic(T_{p+1}, y)$
ALORS ALERT(absorption,x')

Il se peut aussi que deux thématiques de poids équivalents (faible ou fort) fusionnent entre elles, des alertes similaires seront alors créés sur la même base que celles vues précédemment.

Eclatement De la même manière, nous pouvons constater un éclatement par rejet. Cette alerte revient à détecter une sous-thématique (avec un poids assez faible) qui se détache d'une thématique plus importante. Nous la formalisons comme ceci :

SI $split_p(x, y)$ ET $split_p(x, y')$ ET $important\_Topic(T_p, x)$ ET $important\_Topic(T_{p+1}, y)$ ET $weak\_Topic(T_{p+1}, y)$
ALORS ALERT(rejection,y')

De même que pour la fusion, nous pouvons définir deux alertes d'éclatement de thématique de poids équivalents.

Apparition Plutôt que de déceler une thématique de poids assez faible, nous voulons détecter une thématique dont le poids augmente avec les périodes ce qui revient à repérer les thématiques qui apparaissent et qui gagnent en intérêt avec le temps.

SI $appearance_p(\emptyset, y)$ ET $weak\_Topic(T_{p+1}, y)$ ET (Gain_Interest,y)
ALORS ALERT(new_emerging_Topic,y)

Disparition Enfin, la dernière alerte qui nous semble importante et de traquer une thématique jusqu'à sa disparition. Cela revient à déceler une perte progressive d'intérêt suivi d'une disparition totale de la thématique.

SI (Lost_Interest,x) ET $disappearance_p(y, \emptyset)$ ET
ALORS ALERT($Disappearance\_weak\_Topic, x$)

4 Expérimentation

Pour tester le cadre formel présenté ci-dessus, nous avons utilisé un jeu de données contenant des thématiques extraites grâce au logiciel AGAPE (Velcin et Ganascia (2007)). Le premier fichier concerne 461 thématiques (après suppression de 8 thématiques contenant un seul article) datant de la semaine du 11 Décembre 2006. Le second fichier date de la première semaine de Mars 2007 et contient 737 thématiques. Le nombre important de thématique peut être vu comme des *micro-cluster* (Aggarwal et al. (2003)).

Le programme se divise en plusieurs étapes. La première récupère du fichier de sortie d'AGAPE les données nécessaires, à savoir le numéro de la thématique, le nombre total de mots et le nombre d'articles présents dans chaque thématique, les vingt mots les plus fréquents ainsi que le nombre d'occurences associées.

Puis, pour chaque thématique de chaque période, le système va calculer la fréquence de chaque mot : $TF_{w,t,p} = \frac{Nb\_Occur_{w,t,p}}{Nb\_Mot_{t,p}}$ où $TF_{w,t,p}$ représente la fréquence du mot w dans la thématique t à la période p, $Nb\_Occur_{w,t,p}$ représente le nombre d'occurences du mot w dans la thématique t à la période p et $Nb\_Mot_{t,p}$ correspond au nombre total de mots de la thématiques t à la période p. Dans cette expérimentation, nous utilisons seulement la fréquence des termes (TF) et non la mesure TF-IDF car le logiciel d'extraction des thématiques (AGAPE) ne permet pas à un mot de se retrouver dans deux thématiques différentes.

Une fois, les vecteurs pondérés pour T_p et T_{p+1}, le système calcule une matrice des distances (grâce à la mesure du cosinus) entre chaque thématique appartenant à la première période et chaque thématique appartenant à la seconde (Schultz et Liberman (1999)) :

$$\forall x \in T_p \text{ et } \forall y \in T_{p+1}, \text{ d(x,y) = cos(x,y) = } \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Une fois la matrice des distances calculées, le système va extraire P, la partie du produit cartésien $T_p \times T_{p+1}$ tel que : $P = \{(x, y) \mid \ (x \in T_p) \wedge (y \in T_{p+1}) \wedge cos(x, y) > \alpha\}$, α étant défini par l'utilisateur.

Puisque P contient tous les couples (x,y) qui ont une similarité supérieure à α (ici, la similarité est calculée grâce à la distance du cosinus), le système n'a plus qu'à détecter les différentes évolutions possibles à partir de P :

- S'il existe dans P au moins deux x distincts (x,x' $\in T_p$) qui tendent vers un même y (avec y $\in T_{p+1}$) alors une relation de fusion de x et x' en y est détectée.
- S'il existe dans P x $\in T_p$ qui tend vers au moins deux y distincts : y,y' $\in T_{p+1}$ alors une relation d'éclatement de x en y et y' est détectée.
- Enfin, s'il existe dans P un unique x $\in T_p$ qui tend vers un unique y $\in T_{p+1}$ alors une évolution de transformation de x en y est détectée.

Ainsi, nous pouvons déjà voir grâce aux figures 3 et 4 qu'avec une simple distance du cosinus nous arrivons à avoir des résultats encourageants. Le programme a pu déceler des similarités entre les thématiques de la première période et celles de la deuxième. Nous remarquons que la figure 3 qui est le résultat d'une relation d'évolution entre les deux périodes a un cosinus assez important. Cela signifie que les deux thématiques sont assez semblable mais nous remarquons bien que tous les mots ne correspondent pas (13 mots sur 20 sont présents dans les deux thématiques). De même, pour la fusion (figure 4), les deux thématiques de la première période traitent du nucléaire mais sont classées dans deux thématiques séparées car leurs sujets divergent dans la première période. Toutefois, durant la deuxième période ces deux thématiques ne forment plus qu'une.

Cos	Tp	Tp+1
0,7	**Topic 448** = hausse(30), marche(27), points(25), baisse(25), analystes(21), indice(19), taux(17), bourse(17), resultats(16), premier semestre(16), croissance(16), dollars(15), benefice(15), investisseurs(15), trimestre(15), nasdaq(15), dow jones(15), progression(14), marches(14), economie(13) Poids : 0,035	**Topic 674** = hausse(30), analystes(27), indice(27), marche(24), baisse(23), petrole(23),bourse(22), prix(21), fed(21), dow jones(20), echanges(19), new york(18), croissance(17), nasdaq(17), dollars(16), economie(16), surprise(16), immobilier(15), taux(15), investisseurs(15) ; Poids : 0,033

FIG. 3 – *évolution entre deux périodes*

Cos	Tp	Tp+1
0,42	Poids : 0,043 **Topic 447** = onu(32), iran(32), conseil de securite(29), uranium(27), nucleaire(27), enrichissement(25), etats-unis(24), iranien(24), sanctions(23), pays(22), dossier(21), russie(21), allemagne(18), chine(17), affaires etrangeres(17), resolution(17), programme(16), larijani(15), porte-parole(14), refus(14) ;	Poids : 0,036 **Topic 721** = chine(37), **nucleaire**(36), etats-unis(35), russie(34), **japon**(31), pekin(27), **coree du nord**(27), **pourparlers**(25), discussions(24), **pyongyang**(23), **negociations**(21), pays(21), reprise(19), programme(18), **nucleaires**(18) ,sanctions(18), dossier(16), corees(16), americain(15), agence(15) ;
0,3	Poids : 0,016 **Topic 188** = **pyongyang**(13), **japon**(12), missiles(12), **coree du nord**(12), tirs(9), coree du sud(8), **pourparlers**(8), table des **negociations**(8), washington(7), regime communiste(7), test(7), hill(6), multipartites(6), mer(5), taepodong-2(5), portee(5), tournee(5), departement d'etat(5), essai **nucleaire**(5), abc(5) ;	

FIG. 4 – *fusion entre deux périodes*

Cette expérimentation nous a permis de tester le cadre que nous présentons, et plus particuliérement la partie sur le résumé de l'information temporelle par la détection des évolutions. Le système d'alerte étant en cours d'élaboration, nous n'avons pu encore l'intégré.

5 Conclusion et Perspectives

Ce cadre formel est la base d'une réflexion portée sur les problèmes de suivi des données textuelles dans le temps. Il va permettre un résumé visuel des évolutions temporelles des thématiques ainsi qu'une ouverture aux aspects inattendus de l'actualité. En effet, on ne peut pas toujours prédire les événements qui transcendent l'actualité. Que ce soit des cyclones qui dévastent une région, un attentat qui fait des dizaines de morts, certains faits sont, malgré des signes avant-coureurs, difficiles à anticiper, et quand bien même certains l'auraient prévus (des géologues pas exemple), leur représentation dans la presse n'est pas toujours visible au premier coup d'oeil.

Les premiers résultats que nous avons obtenus, nous encouragent à poursuivre dans cette voie. Pour compléter l'évaluation, nous allons faire appel à des professionnels du domaine et plus précisément à l'équipe GSPR [2] de l'EHESS [3] dirigé par Francis Chateauraynaud. Cette équipe travaille sur des questions de controverse dans l'actualité et sur des problèmes de suivi de ces dernières.

Enfin, ce cadre permet de rester à un niveau d'abstraction assez élevé pour tester différentes techniques que ce soit au niveau du regroupement des articles (Agape, K-means, OKM (Cleuziou (2007)), bisecting K-means (Steinbach et al. (2000)), etc.), au niveau du suivi temporel, mais aussi en ce qui concerne les données elles-mêmes (articles de presse, blogs, etc.). Le but est de créer à terme un système assez général, stable et ouvert pour permettre l'intégration de nouvelles données, différentes tailles de périodes, des langues diverses, une modification facile des paramètres, etc.

Cette formalisation va donner la possibilité d'identifier certains événements, de rendre compte de l'actualité de manière synthétique et temporelle en proposant un résumé et un système d'alerte avec l'utilisateur au centre du processus.

Références

Aggarwal, C., J. Han, J. Wang, et P. Yu (2003). A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pp. 81–92. VLDB Endowment.

[2] Groupe de Sociologie Pragmatique et Réflexive

[3] Ecole des Hautes Etudes en Sciences-Sociales

Allan, J. (2002). *Topic Detection and Tracking : Event-Based Information Organization*. Kluwer Academic Publishers.

Allan, J., J. Carbonell, G. Doddington, J. Yamron, et Y. Yang (1998). Topic detection and tracking pilot study : Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Volume 1998. Morgan Kaufmann Publishers, Inc.

Binsztok, H. et P. Gallinari (2002). Un algorithme en ligne pour la détection de nouveauté dans un flux de documents. *Journées Internationales d'Analyse Statistique des Données Textuelles, JADT*.

Cleuziou, G. (2007). OKM : une extension des k-moyennes pour la recherche de classes recouvrantes.

Dumas, P. et Y. Bertacchini (2001). Veille et management des compétences : application de la démarche de veille aux métiers numériques.

Jacquenet, F. et C. Largeron (2005). Extraction automatique d'information inattendue à partir de textes. *Revue des Nouvelles Technologies de l'Information RNTI E 4*.

Jacquenet, F., C. Largeron, et S. Chapaux (2004). Veille Technologique assistee par la Fouille de Textes.

Jakobiak, F. (1991). *Pratique de la veille technologique*. Les éditions d'organisation.

Kontostathis, A., L. Galitsky, W. Pottenger, S. Roy, et D. Phelps (2003). A Survey of Emerging Trend Detection in Textual Data Mining. *Survey of Text Mining : Clustering, Classification, and Retrieval*.

Mei, Q. et C. Zhai (2005). Discovering evolutionary theme patterns from text : an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 198–207. ACM New York, NY, USA.

Schultz, J. et M. Liberman (1999). Topic detection and tracking using idf-weighted cosine coefficient. In *Proceedings of the DARPA Broadcast News Workshop*, pp. 189–192. San Francisco : Morgan Kaufmann.

Steinbach, M., G. Karypis, et V. Kumar (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining*, Volume 34, pp. 35.

Velcin, J. et J. Ganascia (2007). Topic Extraction with AGAPE. *LECTURE NOTES IN COMPUTER SCIENCE 4632*, 377.

Summary

News represent a huge quantity of textual data with a temporally registration at the event level or more widely at the topic level (which groups several news discussing about a same event). This paper begin a reflexion about tracking textually data during time. We present a formal border (at a high abstraction level) that allows topic tracking during time and be sensitive to the unexpected things that can be hided in the data. The aim of this formal border is to detect the news topic evolutions to have, at the end, a temporal resume and a performing alert system to prevent the analyst when something occurs in the news.

Un cadre pour la représentation et l'analyse de débats sur le Web

Anna Stavrianou*, Julien Velcin**, Jean-Hugues Chauchat***

ERIC Laboratoire - Université Lumière Lyon 2, Université de Lyon,
5 avenue Pierre Mendès-France 69676 Bron Cedex, France
* anna.stavrianou@univ-lyon2.fr
** julien.velcin@univ-lyon2.fr
*** jean-hugues.chauchat@univ-lyon2.fr

Résumé. Les débats en ligne sont souvent modélisés par des réseaux sociaux d'utilisateurs représentés sous forme de graphes, chaque noeud correspondant à l'un des intervenants du débat. Ici, nous proposons un nouveau modèle basé sur un graphe de messages : chaque noeud du graphe correspond à l'un des messages échangés et chaque arc relie un message à celui auquel il répond ; ces arcs peuvent être caractérisés par les mots-clés résumant la relation entre les messages connectés. Cette modélisation permet une meilleure représentation de la dynamique du débat ainsi que l'identification de chaînes de discussion. Nous comparons les deux représentations, graphes représentant les utilisateurs et graphes représentant les messages, puis nous analysons la connaissance qui peut être extraite à partir de chacun d'eux. Nos expériences sur des débats réels valident le modèle proposé et montrent les informations complémentaires qui sont apportées par le graphe des messages.

1 Introduction

Le développement du Web2.0 a provoqué la création d'un grand nombre de blogs, de forums et de discussions en ligne. L'analyse de ce type de discussions est très intéressante, tant pour des organismes publics que pour l'industrie ou le secteur commercial. Les discussions en ligne contiennent les intérêts et les avis des internautes, des critiques de produits, la présentation d'idées politiques, etc. En conséquence analyser ces données devient une problématique stratégique.

La fouille des données dans les discussions en ligne exige une représentation appropriée de la discussion. Une discussion en ligne peut être représentée comme un graphe où les noeuds sont des entités (utilisateurs, messages etc.) et les arcs les relations entre les entités. Par conséquent, une discussion peut être analysée par les techniques d'analyse de réseaux sociaux, permettant la cartographie des relations entre personnes, organisations et autres entités qui contiennent des informations. Actuellement la plupart des discussions en ligne sont modélisées par un réseau social des utilisateurs sous forme de graphes représentant les participants de la discussion.

Dans cet article, nous proposons un nouveau modèle pour la modélisation et l'analyse des discussions qui sont menées sur le Web. Chaque discussion a un sujet spécifique, par exemple : « Que pensez-vous de l'appareil photo A ? » ou bien « Faut-il construire de nouvelles centrales nucléaires ? ». Au cours de la discussion, les utilisateurs s'identifient avec un nom d'utilisateur et ils peuvent participer en envoyant des messages. Ils peuvent écrire un nouveau message ou répondre à un message existant. Nous supposons ici que les relations « tel message répond à tel message » sont connues.

Notre contribution est double :

1. la discussion est modélisée par un graphe dont les noeuds représentent les messages échangés plutôt que les utilisateurs qui ont participé à la discussion ;
2. la connaissance relative au contexte est extraite par les techniques de la fouille de textes et elle est utilisée pour caractériser la relation entre les messages connectés.

La représentation proposée permet l'identification directe des élements de la discussion qui sont les plus intéressantes ou qui ont provoqué le début d'un dialogue entre les participants. Aussi, elle peut faciliter l'extraction du résumé de la discussion. Le modèle qui est proposé dans cet article est complémentaire de celui des graphes utilisateurs. Il ne vise pas à les remplacer. Il fournit des informations supplémentaires et il peut être employé en même temps que le réseau social afin d'enrichir les informations extraites d'une discussion.

Ce document est structuré comme suit. La section 2 présente l'état actuel de la recherche. La section 3 présente l'approche suivie pour la représentation des discussions et la section 4 nous permet de valider le modèle proposé par quelques expériences sur des discussions réelles extraites du Web. Enfin, la section 5 permet de conclure et donne des perspectives complémentaires.

2 Etat actuel de la recherche

Un travail récent sur l'analyse de forums est celui de Zhang et al. (2007). Les méthodes d'analyse des réseaux sociaux y ont été utilisées afin de modéliser le forum de Sun (Java) et d'identifier l'expertise des utilisateurs. Dans la représentation graphique, les noeuds représentent des utilisateurs et non des messages comme dans notre approche. De plus, l'objectif de Zhang et al. (2007) est de trouver des experts ; il est différent du notre.

Un travail qui sépare un ensemble d'utilisateurs d'un *newsgroup* entre les partisans du pour et ceux du contre est présenté dans Agrawal et al. (2003). Dans ce travail, le contenu de chaque texte n'est pas analysé parce que les auteurs considèrent que les méthodes statistiques ne fonctionnent pas pour de petits messages. Contrairement à eux, nous considérons le contenu du texte et nous l'analysons afin d'identifier des similarités dans le discours.

Dans Fisher et al. (2006), la discussion est représentée par le reseau social des utilisateurs. L'objectif est l'interpretation d'une communauté en ligne afin d'identifier les rôles de chaque utilisateur. Les rôles, c'est à dire la relation d'un noeud du reseau avec ses voisins a été présenté aussi dans Scripps et al. (2007).

3 Modèle d'une discussion

Afin de représenter la structure d'une discussion, nous employons une représentation basée sur les graphes. Les noeuds du graphe représentent les entités et les arcs les relations entre ces entités. Dans notre problématique, nous identifions deux types d'entités : les utilisateurs qui participent à la discussion et les messages écrits par les utilisateurs. Les utilisateurs et les messages sont des entités qui sont connectées par des relations « répondre-à ».

A notre connaissance, la plupart des recherches existantes considèrent les *utilisateurs* comme les noeuds du graphe. Dans cet article, nous utilisons les *messages* comme les entités principales représentées par les noeuds du graphe et nous analysons l'information qui peut être extraite à partir de ce modèle.

Par conséquent, nous représentons les discussions en ligne par un graphe orienté G=(X,U), où X est l'ensemble des noeuds qui représentent des objets de type message et U sont les arcs qui représentent les relations « répondre-à ». Les objets de type message ont le format M=(m,a), où « m » représente le contenu du message et « a » son auteur. De cette façon, les informations sur l'auteur ne sont pas perdues. Les arcs sont marqués par les mots-clés inclus dans le message de réponse qui sont extraits et filtrés par les techniques de la fouille de textes (Stavrianou et al. (2007)). Nous appelons ce graphe : « graphe des messages ».

Le modèle proposé, représenté par les graphes de messages, offre davantage d'informations que celui qui est uniquement basé sur les utilisateurs. Il permet l'extraction des chaînes de discussion, c'est à dire l'ensemble des messages qui sont connnectés entre eux et qui descendent d'un message racine du graphe.

Souvent dans les discussions en ligne apparaissent des messages qui ne donnent aucune information (ex : « Bonjour. », « lol », « oui, je suis là. »). Identifier des mots-clés pour caractériser les arcs du graphe rend plus facile l'identification de la partie de la discussion qui contient les informations les plus intéressantes. Ceci permet, par exemple, de concentrer l'analyse sur un sous-ensemble de messages plutôt que sur tous les messages de la discussion. De plus, la présence des mots-clés dans le modèle peut faciliter l'extraction d'un résumé de la discussion.

Avec le graphe des messages, on peut obtenir (liste non exhaustive) :

1. les chaînes de discussion qui montrent quels utilisateurs parlent des mêmes sujets ;
2. les chaînes de discussion qui sont les plus intéressantes ;
3. les messages qui ont créé beaucoup de réactions ;
4. l'ensemble des réponses données à un message spécifique ;
5. un résumé du contenu de la discussion.

Ces informations ne sont pas directement accessibles quand nous représentons une discussion par un réseau social d'utilisateurs ; en effet, dans le *graphe d'utilisateurs*, les relations « répondre-à » n'impliquent pas toujours une similitude de sujet thématique puisque deux utilisateurs peuvent avoir répondu entre eux quatre fois dans quatre chaînes différentes. Au contraire, dans un *graphe des messages* qui se concentre sur les messages échangés, il est possible de calculer des similarités de contenu sur des messages apparaissant dans les mêmes chaînes de discussion.

Le graphe des messages identifie également les messages-clés c'est à dire ceux qui ont créé le plus de réactions ou qui ont provoqué un dialogue. L'analyse est concentrée sur le contenu plutôt que sur les participants de la discussion. Un message qui a reçu beaucoup de réponses est probablement plus intéressant qu'un message qui n'en a pas reçu. Avec cette

représentation, nous pouvons rapidement identifier le nombre de réponses qui ont été données à un message spécifique aussi bien que les réponses elles-mêmes.

Les différences entre le modèle proposé et celui généralement employé sont récapitulées dans le tableau 1.

	Graphe d'utilisateurs	Graphe de messages
Interaction	On peut observer l'interaction entre les utilisateurs.	On peut observer comment les messages construisent des chaînes de discussion.
Popularité	Si plusieurs arcs descendent d'un noeud utilisateur, alors cet *utilisateur* est intéressant parce que il a reçu beaucoup de messages.	Si plusieurs arcs descendent d'un noeud de message, alors ce *message* est intéressant parce qu'il a provoqué beaucoup de réactions.
Communauté	Une communauté d'utilisateurs montre les utilisateurs qui parlent entre eux (amitié, intérêt sur le même sujet, accord, désaccord).	Une communauté de messages montre les réactions en chaîne et des relations de similarité entre les contenus.
Chaînes de discussion	Impossible	Identifier les chaînes de discussion dans le débat.
Contenu	Impossible	Approximation du contenu de la discussion.
Transitivité (A→ B→C)	On ne peut pas assumer que l'utilisateur A discute indirectement avec l'utilisateur C car le sujet peut être différent.	Les liens entre le message A et le message C montrent la possibilité de similarité de contenu.
Taille du graphe	Grandit proportionnellement au nombre de participants.	Grandit proportionnellement au nombre de messages.

TAB. 1 – *Différences entre le graphe des utilisateurs et le graphe des messages.*

Nous pouvons voir dans le tableau 1 que l'inconvénient d'une représentation basée sur les messages est la taille du graphe qui peut devenir très grande pour les discussions longues contenant beaucoup de noeuds de messages. Concernant ces discussions, des méthodes de visualisation de grands graphes ont été développées ; les données textuelles pourraient également être contrôlées en employant des techniques de résumé de textes ou de classification (Sebastiani (2002)).

Le graphe utilisateurs et le graphe des messages ont des objectifs différents et extraient des informations complementaires. Les deux permettent de donner une structure à la discussion et facilitent l'analyse de cette discussion en extrayant l'information utile à partir de la représentation structurée. Dans le modèle de l'analyse de discussions, l'utilisation combinée de ces deux types de graphes est préférable.

4 Expérimentations

Afin de valider notre modèle de discussions, nous avons fait des expérimentations avec des discussions courtes qui se composent de 6 à 24 messages où les relations «répondre-à» sont connues. Dans un avenir proche, nous prévoyons d'analyser de plus grands ensembles de données. Les expériences présentées dans cet article ont été effectuées avec des données issues de discussions de sites Web en anglais. De la même manière, nous pouvons extraire de l'information et représenter des discussions en ligne en langue française ou dans n'importe quelle autre langue.

L'approche que nous suivons se compose de trois étapes. Tout d'abord, nous identifions la structure de chaque discussion et nous la représentons par un graphe. Ensuite nous extrayons l'information relative aux mots-clés à partir des messages de la discussion, et finalement nous marquons le graphe. Pour des raisons expérimentales, nous avons développé une interface pour la visualisation des deux graphes. L'implémentation est faite en utilisant la bibliothèque java JUNG (http://jung.sourceforge.net).

Notre ensemble de données se compose de 20 débats au total qui apparaissent dans la « Digital Camera Magazine Community » (http://community.dcmag.co.uk) concernant des discussions autour de la photographie ou dans des groupes de Google (groups.google.com) concernant des destinations de vacances. Pour chaque discussion, nous avons identifié la structure en créant un graphe dont les noeuds montrent les messages échangés et les arcs précisent les relations « répondre-à ».

Une fois que l'extraction des mots-clés est faite, nous avons combiné cette information avec la structure de la discussion. En représentant la discussion comme graphe des messages nous pouvons facilement observer les sujets qui apparaissent dans la discussion en suivant les différentes chaînes dans le graphe. L'information de mot-clés placée sur les arcs permet l'identification des messages potentiellement « intéressants ».

Le tableau 2, donne le nombre de messages échangés et le nombre des participants pour les 20 débats.

Débat	Messages	Utilisateurs
1	24	13
2	19	12
3	8	6
4	16	11
5	7	5
6	11	7
7	13	4
8	16	6
9	19	8
10	18	6
11	12	6
12	10	9
13	12	10
14	15	7
15	11	4
16	14	10
17	15	9
18	7	7
19	6	6
20	13	13

TAB. 2 – *Statistique sur 20 débats différents.*

Le nombre de messages dans les trois dernières discussions est égale au nombre de participants, ce qui signifie que chaque message a été écrit par un utilisateur différent. Il n'y a donc pas de réel dialogue entre deux utilisateurs. Par contre, dans la première discussion les utilisateurs ont participé en envoyant 2 messages en moyenne.

4.1 Analyse détaillée d'un exemple

Nous développons le débat no 5 pour illustrer la différence d'information fournie par un graphe d'utilisateurs et un graphe des messages. La discussion est en anglais, le titre est « Anyone ever dealt with Y ? » (Y représente ici le nom d'un site d'achats en ligne), 5 utilisateurs A, B, C, D, E y ont participés et 7 messages y sont échangés. L'échange des messages est montré dans le tableau 3.

Dans le tableau 3, la colonne « auteur » montre l'auteur du message et la colonne « Répondre-à » montre à qui et à quel message la réponse se réfère. La colonne « Messages » contient les messages. Les mots-clés qui sont extraits de chaque message sont en italiques.

La représentation de cette discussion courte par un graphe des messages est illustrée sur le schéma 1. Pour des raisons de lisibilité les mots-clés ne sont pas ajoutés sur le schéma. La discussion a trois chaînes : {(1,A), (2,B), (3,A)}, {(1,A), (4,C)}, {(1,A), (5,D), (6,A), (7,E)} dont la racine est le message initial. Le graphe des utilisateurs est montré sur le schéma 2.

Message	Auteur	Répondre-à	Messages
1	A	-	I'm *thinking* of *buying* from these people, ever *heard* of them ? Any *recommendations/warnings* would be gratefully *received*.
2	B	B→ A(message 1)	Yes chap. I *bought* my S70mm *macro lens* from them and a macro *flash*. I had no *problems* at all.
3	A	A→ B(message 2)	*Thanks fella*, *nice* to *know*. Now to *spend* some *cash* !
4	C	C→ A(message 1)	I *bought* some H *filters* from them at the end of *last year*, at the same time their *prices* were very *competitive*. *Delivery* was pretty *quick* if I *remember* right and I certainly had no *problems* of any sort. If the *price* was *right* I would *use* them again.
5	D	D→ A(message 1)	yes, had very *good service* from them. *good prices* too.
6	A	A→D(message 5)	*Thanks guys*, they are the *cheapest* I can find for S70 *f2.8*.
7	E	E→A(message 6)	I *bought* some H *filters* from them *last year*, *delivery* was *slow* but I did *get* a *courtesy* e-mail *apologising* and *explaining* probable *delivery dates* which were then *met*. Overall I would definitely *buy* from them again if they were the most *competitive*.

TAB. 3 – *Messages de l'exemple.*

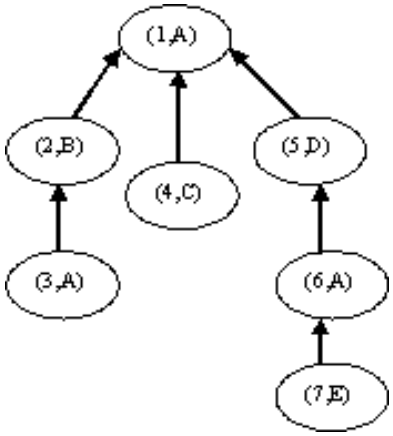


FIG. 1 – *Graphe des messages de l'exemple.*

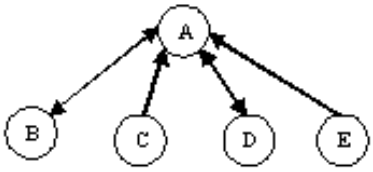


FIG. 2 – *Graphe des utilisateurs de l'exemple.*

Les graphes sur les schémas 1 et 2 représentent des informations différentes pour une même discussion. Le graphe des utilisateurs montre l'interaction entre les participants de la discussion. Nous ne savons ni qui a lancé la discussion ni son contenu. Le graphe des messages suit la structure temporelle de la discussion. Regardant le graphe de messages, nous pouvons voir que la discussion entière est relative au sujet du premier message. Dans cet exemple en particulier, tous les internautes ont répondu à la question posée dans le premier message.

En utilisant les mots-clés des messages nous pouvons comprendre le sujet général de la discussion et nous pouvons identifier rapidement les messages dont le contenu est intéressant ou non. Dans notre exemple, en regardant les mots-clés et sans lire la discussion nous comprenons que les internautes sont en général satisfaits par le service et les prix de Y. La chaîne de discussion {(1,A), (2,B), (3,A)} représente un dialogue complet dans le sens où le message 2 répond à la question posée dans le message 1. Le message 3 écrit par l'auteur du message 1 remercie de la réponse du message 2.

4.2 Discussion

L'expérimentation avec des discussions réelles nous a permis d'observer les caractéristiques de discussions courtes et d'identifier comment les utilisateurs se comportent au cours de telles discussions. Nous avons noté qu'au cours des discussions courtes ou celles avec peu d'utilisateurs, chacun est relié avec tous les autres. Les utilisateurs appartiennent à la même communauté et souvent cette communauté est une clique. En conséquence, si nous représentons la discussion du point de vue des utilisateurs, la notion des cliques et des communautés entre les utilisateurs peut ne pas sembler intéressante. Par contre, la notion de communautés entre les *messages* a un sens afin d'identifier des chaînes de discussion. La découverte et l'analyse des communautés et des rôles (relation d'un noeud à ses voisins) au cours des discussions en ligne est un domaine de recherche récent (Du et al. (2007), Fisher et al. (2006), Scripps et al. (2007), Zhou et al. (2007)). Au cours d'une discussion représentée par un graphe des messages, connaître les rôles des messages et leur contenu (ou leur résumé) pourrait aider chaque participant à découvrir plus vite comment il peut participer à la discussion ou à qui parler en premier lieu, au lieu de perdre du temps à lire tous les messages.

Les expériences ont confirmé que les graphes des messages contiennent des informations qui ne sont pas montrées par les graphes des utilisateurs, comme par exemple le déroulement de la discussion. La représentation d'une discussion avec un graphe utilisateurs nous permet d'identifier qui parle à qui ou qui pourrait être considéré en tant qu'expert en matière de discussion (Zhang et al. (2007)). De son côté, un graphe des messages nous permet de voir comment la discussion évolue, extrait les différentes chaînes et identifie rapidement les parties des discussions qui sont intéressantes à analyser.

Ceci permet la fouille et la sélection des données les plus pertinentes. Si nous avions un graphe des utilisateurs où les arcs avaient l'information de mot-clés, nous ne connaîtrions pas la chaîne de discussion dans laquelle un dialogue intéressant a eu lieu ; deux utilisateurs peuvent échanger beaucoup de messages non informatifs (formule de politesse) avant de commencer à parler sur le fond, et en suite il peuvent se répondre à des moments différents sur des sujets différents.

5 Conclusion et perspectives

Dans cet article, nous avons proposé un nouveau modèle qui représente les discussions en ligne par des graphes de messages. Ceci permet une représentation orientée par le contenu et la dynamique d'une discussion. Le nouveau modèle permet l'identification des parties de la discussion les plus intéressantes et facilite la classification de la discussion par rapport aux thématiques qui ont été abordées.

Puisque les graphes de messages n'ont pas été explorés jusqu'ici, les travaux de recherche future sont nombreux. Nous travaillons actuellement à enrichir le graphe avec des polarités d'opinion pour chaque message. Ceci nécessite évidemment l'application des méthodologies de la fouille des données d'opinion (Hatzivassiloglou et Mckeown (1997), Hu et Liu (2004), Pang et al. (2002), Turney et Littman (2003)). L'extraction de mots-clés avec la combinaison d'information d'opinion permettra l'identification de l'accord ou du désaccord entre les utilisateurs. De plus, nous finalisons une plateforme de discussions en ligne qui nous permettra de réaliser les expérimentations à plus grande échelle et d'améliorer le traitement des textes.

Dans ce travail les liens entre les messages (quel répond à quel) sont connues, mais souvent ces liens sont manquants. Nous envisageons de travailler sur la recherche automatique des liens entre les messages qui apparaissent dans une discussion. Dans ce cas, les methodes de la fouille de textes sont indispensables. Un exemple est l'extraction de phrases-clès qui est presenté dans Hammouda et al. (2005).

De nouvelles mesures sur les graphes de messages doivent être explorées et les mesures existantes basées sur les réseaux sociaux doivent être adaptées aux nouveaux graphes. Parmi celles existantes, on trouve la recherche de la popularité des noeuds, la participation des noeuds dans les communautés, la centralité des noeuds, etc. (Carrington et al. (2005)). Ces algorithmes sont une zone de recherche dans les graphes des messages.

Une autre piste pour le futur est de combiner les deux type de graphes (graphes des des utilisateurs et graphes des messages) afin d'analyser une discussion. Par exemple, nous pouvons utiliser le graphe des utilisateurs pour extraire les utilisateurs qui sont les experts (Zhang et al. (2007)) dans le domain de la discussion. Aprés en utilisant cet information nous pouvons extraire du graph des messages les chaînes de discussion où les experts qui ont participé.

Références

Agrawal, R., S. Rajagopalan, R. Srikant, et Y. Xu (2003). Mining newsgroups using networks arising from social behavior. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*.

Carrington, P., J. Scott, et S. Wasserman (2005). *Models and Methods in Social Network Analysis*. New York: Cambridge University Press.

Du, N., B. Wu, X. Pei, B. Wang, et L. Xu (2007). Community detection in large-scale social networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 16–25. ACM.

Fisher, D., M. Smith, et H. Welser (2006). You are who you talk to: Detecting roles in usenet newsgroups. In *HICSS '06: Proceedings of the 39th Annual Hawaii International Conference on System Sciences*. IEEE Computer Society.

Hammouda, K., D. Matute, et M. Kamel (2005). Corephrase: Keyphrase extraction for document clustering. In *MLDM: Machine Learning and Data Mining in Pattern Recognition*, pp. 265–274.

Hatzivassiloglou, V. et K. Mckeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics*, pp. 174–181.

Hu, M. et B. Liu (2004). Mining and summarizing customer reviews. pp. 168–177. ACM.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 79–86.

Scripps, J., P.-N. Tan, et A.-H. Esfahanian (2007). Node roles and community structure in networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 26–35. ACM.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys 34*(1), 1–47.

Stavrianou, A., P. Andritsos, et N. Nicoloyannis (2007). Overview and semantic issues of text mining. *SIGMOD Record 36*(3), 23–34.

Turney, P. et M. Littman (2003). Measuring praise and criticism: inference of semantic orientation from association. *ACM TOIS 21*(4), 315–346.

Zhang, J., M. Ackerman, et L. Adamic (2007). Expertise networks in online communities: Structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pp. 221–230.

Zhou, D., I. Councill, H. Zha, et C.-L. Giles (2007). Discovering temporal communities from social network documents. In *ICDM: International Conference on Data Mining*, pp. 745–750. IEEE Computer Society.

Summary

The online discussions are often modeled as a social network of users and they are represented by a graph where each node denotes a participant of the discussion. In this paper, we propose a new framework for discussion analysis. It is based on a graph of messages: each node corresponds to a message of the discussion and each edge (directed) points out which message the specific node replies to. The edges can be weighted by the keywords that characterize the connected messages. This model allows a better representation of the content of the discussion and facilitates the identification of discussion chains. We compare the two representations: the user-based and the message-based graph and we analyze the different information that can be extracted from them. Our experiments with real data validate the proposed framework and show the additional information that can be extracted from a message-based graph.