Analyse de gazouillis en ligne

Sandra Bringay [*,**] , Nicolas Béchet [***], Flavien Bouillot [*], Pascal Poncelet [*],
Mathieu Roche [*], Maguelonne Teisseire [****]

[*]LIRMM – CNRS, 161 rue Ada, 34095 Montpellier, France
{bringay,bouillot,poncelet,mroche}@lirmm.fr
[**]Univ. Montpellier 3
[***] INRIA Rocquencourt- Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France
nicolas.bechet@inria.fr
[****] Cemagref, UMR TETIS, 500 rue Jean-François Breton, 34093 Montpellier, France
maguelonne.teisseire@cemagref.fr

Résumé. Les tweets échangés sur Internet constituent une source d'information importante même si leurs caractéristiques les rendent difficiles à analyser (140 caractères au maximum, notations abrégées, ...). Dans cet article, nous définissons un modèle d'entrepôt de données permettant de valoriser et d'analyser de gros volumes de tweets en proposant des mesures pertinentes dans un contexte de découverte de connaissances. L'utilisation des entrepôts de données comme outil de stockage et d'analyse de documents textuels n'est pas nouvelle mais les mesures ne sont pas adaptées aux spécificités des données manipulées. Les résultats des expérimentations sur des données réelles soulignent la pertinence de notre proposition.

1 Introduction

Ces dernières années, le développement du web social et collaboratif 2.0 a rendu les internautes plus actifs au sein des réseaux participatifs. Les blogs pour diffuser son journal intime de manière massive, les dépêches RSS pour suivre les dernières informations sur un thème, les tweets pour publier ses faits et gestes, sont désormais extrêmement répandus. Simples à créer et gérer, ces outils sont utilisés par les internautes, les entreprises ou autres organisations pour communiquer à propos d'eux-mêmes ou des phénomènes qui les concernent. L'exploitation de ces nouvelles formes de publications s'inscrit dans une logique d'intelligence collective et suscite des applications inattendues en terme d'aide à la décision. En effet, les décideurs peuvent utiliser ces gros volumes d'information comme nouvelles ressources documentaires pour y puiser automatiquement de l'information.

Depuis son apparition en 2006, le site Twitter [1] s'est développé de telle manière qu'il est actuellement le $10^{ième}$ site le plus visité au monde [2]. Twitter est une plate-forme de "micro-blogging". Il s'agit d'un système de partage d'informations via lequel les utilisateurs ont la

1. http ://twitter.com
2. http ://www.alexa.com/siteinfo/twitter.com

possibilité de suivre d'autres utilisateurs qui postent des messages courts ou de diffuser ses messages à tous ses abonnés. En janvier 2010, le nombre de tweets échangés a atteint 1,2 milliards et plus de 40 millions de tweets sont échangés par jour en moyenne [3]. Lorsqu'un utilisateur "suit" une personne, il reçoit tous les messages qu'elle poste, et inversement, lorsque que cet utilisateur "tweete", tous ses "suiveurs" reçoivent ses messages. Les tweets sont associés à des méta-informations qui peuvent être non-incluses dans les messages (la date, la géolocalisation...) ou bien incluses dans le message sous la forme de tags (balises) ayant une signification particulière : par exemple lorsque le tag @nomutilisateur apparait dans un tweet, cela signifie que l'on s'adresse à un utilisateur particulier ; #sujet (hastag) attribue un sujet au message ; RT signifie que le message a été re-tweeté, c'est-à-dire récupéré dans les messages postés par un utilisateur que l'on suit pour qu'il soit visible par ses propres suiveurs. Les tweets peuvent donc être représentés de manière multidimensionnelle en prenant en compte l'ensemble des ces méta-informations et des relations temporelles associées.

Des applications récentes ont été proposées pour analyser l'information à partir des gros volumes de tweets produits au cours du temps, comme par exemple, le suivi de tendances, le repérage de buzz... Toutefois, il n'existe pas à notre connaissance d'approche exploitant leurs caractéristiques multidimensionnelles. Dans cet article, nous nous intéressons plus précisément aux entrepôts de données, introduits au début des années 1990 (Codd et al. (1993)) comme outil de stockage et d'analyse de données multidimensionnelles et historisées. Il devient ainsi possible de manipuler un ensemble d'indicateurs (ou mesures) en fonction de différentes dimensions éventuellement munies d'une ou plusieurs hiérarchies. Les opérateurs associés permettent une navigation intuitive selon différents niveaux de hiérarchies.

Dans cet article, nous nous intéressons à définir différents opérateurs de manipulation afin d'identifier des tendances, rechercher les top-k mots (ou groupe de mots) les plus significatifs sur une période de temps, les plus représentatifs d'une ville ou d'un pays, d'un certain mois, d'une certaine année... et l'impact des hiérarchies sur ces opérateurs. Nous proposons ainsi une mesure adaptée, appelée "TF-IDF adaptatif", qui permet d'identifier les mots les plus significatifs selon le niveau des hiérarchies du cube (e.g. via la dimension localisation). L'utilisation des hiérarchies est propagée aux mots des tweets eux-même nous permettant de proposer une contextualisation afin de mieux en appréhender le contenu. L'illustration proposée concerne les évolutions des maladies (e.g. en novembre et décembre, la plupart des tweets concernaient la grippe) en utilisant le thesaurus MeSH [4] (Medical Subject Headings) utilisé entre autre pour indexer les articles de PubMed [5].

Dans la suite de cet article, la section 2 décrit un modèle de données adapté à la problématique des cubes de tweets et détaille la mesure proposée. Dans la section 3, nous considérons une hiérarchie sur les mots des tweets et proposons une nouvelle approche pour contextualiser le tweet par rapport à cette hiérarchie. Les premières analyses et résultats sont décrits dans la section 4. Avant de conclure cet article en présentant les travaux futurs, nous proposons un état de l'art du domaine dans la section 5.

3. http ://blog.twitter.com/2010/02/measuring-tweets.html
4. http ://www.ncbi.nlm.nih.gov/mesh
5. http ://www.ncbi.nlm.nih.gov/pubmed/

2 Quelle mesure d'agrégation pour les tweets ?

Dans cette section, nous introduisons le modèle adopté pour spécifier un cube de tweets. Nous considérons, dans ce modèle, qu'il n'existe pas de hiérarchie prédéfinie sur les mots et nous nous focalisons sur la définition d'une nouvelle mesure associée aux différents mots d'un tweet. Notons en effet que la mesure proposée sera utilisée comme fonction d'agrégation. Nous utiliserons ainsi par la suite le terme de "mesure d'agrégation".

2.1 Définition préliminaire

Selon Pérez-Martínez et al. (2008), une table de faits F est définie sur le schéma $D = \{T_1, \ldots, T_n, M\}$ où T_i $(i = 1, .., n)$ correspondent aux dimensions et M correspond à une mesure. Chaque dimension T_i est définie sur un domaine $d = dom(T_i)$ partitionné en un ensemble de catégories (ou niveaux de granularité) C_j. On a donc $D = \cup_j C_j$. D doit être muni d'un ordre partiel $\sqsubseteq_D$ permettant de comparer les valeurs du *domaine* d. Chaque catégorie représente les valeurs associées à un niveau de granularité. Nous notons $e \in D$ pour préciser que e est une valeur de dimension de D, s'il existe une catégorie $C_j \subseteq D$ telle que $e \in \cup_j C_j$. Notons que deux catégories particulières sont distinguées et sont présentes sur toutes les dimensions : $\perp_D$ et $\top_D \in C_D$ correspondant respectivement au niveau de plus fine et de plus forte granularité. Dans le cadre de notre approche, l'ordre partiel défini sur les domaines des dimensions correspond à l'inclusion ensembliste des mots clés associés aux valeurs des dimensions considérées. Ainsi, soient deux valeurs $e_1, e_2 \in \cup_j C_j$, nous avons $e_1 \sqsubseteq_D e_2$ si e_1 est logiquement contenu dans e_2.

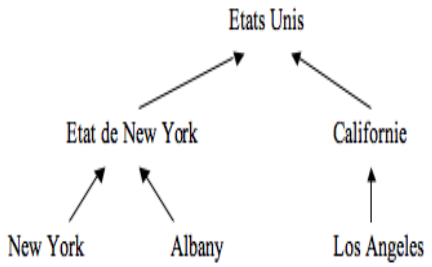

FIG. 1 – *Une hiérarchie associée à la dimension localisation*

Exemple 1 *La dimension Localisation de la figure 1 possède les catégories* $\perp_{localisation} = Ville \leq Région \leq Pays \leq \top_{localisation}$. *Les valeurs de dimensions sont* $dom(Localisation)$ = $\{New\ York, Albany, Los\ Angeles, Etat\ de\ New\ York, Californie, Etats\ Unis, ...\}$ *réparties dans ces catégories (niveaux de granularité) de la manière suivante :* $Ville = \{NewYork, Albany, Los\ Angeles\}$, $Région = \{Etat\ de\ New\ York, Californie\}$, $Pays = \{Etats\ Unis...\}$. *L'ordre partiel* $\sqsubseteq_D$ *sur les valeurs des dimensions peut bien entendu être généralisé aux catégories : pour* $C_1, C_2 \in C_D$, *on a alors* $C_1 \leq_D C_2$ *si* $\exists e_1 \in C_1, e_2 \in C_2$ $tels\ que$ $e_1 \sqsubseteq_D e_2$. *Par exemple, nous avons :* $LosAngeles \sqsubseteq_D Californie \sqsubseteq_D$

$EtatsUnis \sqsubseteq_D \top$. *La prise en compte de hiérarchie dynamique est telle que toutes les catégories de cette dimension doivent respecter l'ordre partiel défini.*

2.2 Le modèle de données

Nous instancions le modèle de données de la section précédente afin de prendre en compte les différentes dimensions de description et bien entendu une dimension spécifique associée aux mots des tweets.

Considérons, par exemple, l'analyse des tweets propres au régime Duncan (e.g. "*The Dukan diet is the best diet ever, FACT ! ! ! Its just meat for me for the next 5 day YEESSS*") et les problématiques de la santé associées. Nous souhaitons, par exemple, suivre l'avis des différentes personnes sur l'efficacité d'un régime. De manière à extraire les tweets correspondant à ce sujet, nous interrogeons Twitter à l'aide d'un ensemble de mots germes associés aux régimes : *duncan, diet, slim, protein, ...* Dans ce cas, à l'origine, les valeurs de la dimension mot sont $dom(mot) =$*{duncan, diet, slim, protein, ...}*.

La figure 2 illustre le modèle de données utilisé. Nous retrouvons la dimension *localisation* présentée dans la section précédente et la dimension *temps* telle que : $\perp_{temps} = jour \leq mois \leq semestre \leq année \leq \top_{temps}$. Le domaine de la dimension *mot* est celui des mots germes enrichi des mots apparaissant fréquemment avec ces derniers. Dans la table de faits, différentes mesures peuvent être utilisées. Traditionnellement celle-ci correspond au TF-IDF. Ce principe est détaillé dans la section suivante.

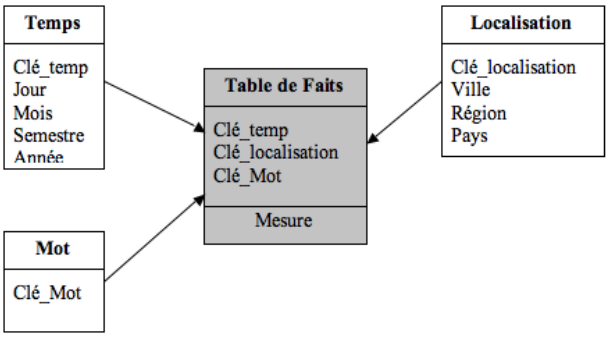

FIG. 2 – *Le schéma associé à l'application régime*

2.3 Vers une mesure adaptée

Le fait de nous appuyer sur les seules connaissances de la hiérarchie dans un cube ne permet pas toujours une agrégation de qualité (i.e. représentant une situation réelle). Par exemple, dans le cadre d'une région, les mots caractéristiques d'un tweet ne sont pas nécessairement les mêmes que pour une ville.

La mesure d'agrégation que nous proposons repose sur des approches issues de la Recherche d'Information (RI). La prise en compte des informations hiérarchiques dans les mesures d'agrégation constitue une des contributions de notre approche.

Dans notre processus, la première étape consiste à fusionner le nombre d'occurrences des mots propres à un niveau. De manière concrète, nous listons tous les mots situés dans les tweets qui correspondent à un niveau donné (e.g. Ville, Région, Pays). Si l'utilisateur souhaite axer sa recherche au niveau de la ville V, les mots des tweets émis dans cette ville forment un vecteur. Nous pouvons appliquer ce même principe au niveau de la Région via une opération de $Roll\text{-}up$.

Le but de nos travaux est de mettre en exergue les mots discriminants par niveau. Dans ce contexte, nous mettons en œuvre une mesure que nous appelons $TF\text{-}IDF$ *adaptatif* qui consiste à ordonner les mots selon le niveau où l'utilisateur se situe. De manière identique aux travaux de Grabs et Schek (2002) qui proposent la mesure *ief* (*inverted element frequency*), nos travaux combinent les informations hiérarchiques avec des mesures statistiques issues du domaine de la Recherche d'Information. Une première version de notre mesure *adaptative* a été proposée dans (Bringay et al. (2010)).
Traditionnellement, la mesure $TF\text{-}IDF$ donne un poids plus important aux mots caractéristiques d'un document (Salton et al. (1975)). Dans, un premier temps, il est alors nécessaire de calculer la fréquence d'un terme (*Term Frequency*). Celle-ci correspond au nombre d'occurrences de ce terme dans le document considéré. Ainsi, pour le document d_j et le terme t_i, la fréquence du terme dans le document est donnée par l'équation suivante :

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

où $n_{i,j}$ est le nombre d'occurrences du terme t_i dans d_j. Le dénominateur correspond au nombre d'occurrences de tous les mots dans le document d_j.

La fréquence inverse de document (*Inverse Document Frequency*) mesure l'importance du terme dans l'ensemble du corpus. Elle consiste à calculer le logarithme de l'inverse de la proportion de documents du corpus qui contiennent le terme. Elle est définie de la manière suivante :

$$IDF_i = log_2 \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

où $|D|$ représente le nombre total de documents dans le corpus et $|\{d_j : t_i \in d_j\}|$ est le nombre de documents dans lesquels le terme t_i apparaît.

Enfin, le poids s'obtient en multipliant les deux mesures :

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i$$

Notons que, dans cet article, notre approche *adaptative* s'appuie sur la mesure classique $TF\text{-}IDF$, mais elle peut être aisément associée à d'autres pondérations (par exemple, Okapi, LTU ou ATC, détaillées respectivement dans Buckley et al. (1995), Robertson et Walker (1999) et Jin et al. (2001)).

Dans notre cas, nous ne calculons pas les mots représentatifs par rapport au nombre de documents mais par rapport au niveau de granularité souhaité. Ainsi, nous définissons le IDF *adaptatif* de la manière suivante :

$$IDF_i^k = log_2 \frac{|E^k|}{|\{e_j^k : t_i \in e_j^k\}|} \quad (1)$$

où $|E^k|$ représente le nombre total d'éléments de type k (dans notre exemple, $k = \{$ *Ville*, *Région*, *Pays* $\}$) qui correspond au niveau de la hiérarchie que le décideur souhaite agréger. $|\{e_j : t_i \in e_j\}|$ est relatif au nombre d'éléments de type k dans lequel le terme t_i apparaît.

Cette mesure permet d'attribuer un poids adapté au niveau d'agrégation décidé par l'utilisateur. Ainsi, nous calculons ce poids $TF\text{-}IDF_i^k$ pour chacun des mots t_i. Nous pouvons ainsi conserver les n mots ayant les poids les plus élevés.

Les résultats obtenus selon ces différents paramètres seront discutés dans la section 4 de cet article.

Exemple 2 *Supposons que dans le contexte de la figure 1, nous ayons le mot "protein" présent une fois dans les tweets de New-York, zéro fois dans les tweets d'Albany et trois fois dans les tweets de Los Angeles.*

Dans cet exemple, le $TF\text{-}IDF$ adaptatif de ce mot au niveau de la $Ville$ (3 villes sont répertoriées) est[6] *:*

- $TF\text{-}IDF^{Ville}_{New-York} = 1 \times log_2 \frac{3}{2} = 0.58$
- $TF\text{-}IDF^{Ville}_{Albany} = 0 \times log_2 \frac{3}{2} = 0$
- $TF\text{-}IDF^{Ville}_{LosAngeles} = 3 \times log_2 \frac{3}{2} = 1.74$

Ceci met en relief que le mot "protein" est typique de la ville de Los Angeles et à moindre mesure de la ville de New-York. Par la suite, si le décideur effectue une opération de Roll-up pour se situer au niveau de la région, nous avons alors :

- $TF\text{-}IDF^{Region}_{EtatdeNew-York} = 1 \times log_2 \frac{2}{2} = 0$
- $TF\text{-}IDF^{Region}_{Californie} = 3 \times log_2 \frac{2}{2} = 0$

Ce mot apparaissant dans les deux Régions (i.e. la valeur de l'IDF adaptatif est 0), il n'est donc pas caractéristique des deux régions répertoriées.

Considérons à présent que le décideur souhaite se focaliser sur un pays. Dans la définition proposée de la mesure $TF\text{-}IDF$ *adaptatif*, l'IDF considère l'ensemble des villes de tous les pays (i.e. E^k) et ne reflète donc pas tous les mots spécifiques à toutes les villes d'un pays. Pour cela, nous proposons une mesure spécifique appelée $TF\text{-}IDF$ *adaptatif-local* pour laquelle le numérateur ne considère que le nombre de fils du niveau considéré.

3 Une hiérarchie sur les mots comme outil de contextualisation des tweets

Dans cette section, nous adoptons une hiérarchie sur les mots pour permettre la contextualisation des tweets.

3.1 Le modèle et les données

Pour la hiérarchie sur les mots, nous utilisons le thesaurus MeSH (Medical Subject Headings)[7] de la National Library of Medicine aux Etats Unis qui est composé d'un ensemble

6. Dans cet exemple, pour plus de visibilité, le nombre d'occurrences n'a pas été normalisé contrairement aux expérimentations présentées en section 4

7. http ://www.nlm.nih.gov/pubs/factsheets/mesh.html

de mots correspondant à des concepts du domaine. Ces derniers sont associés à une structure hiérarchique de 12 niveaux pour permettre la recherche à différents niveaux de la hiérarchie. En 2011, 26 142 concepts sont disponibles dans MeSH. Au niveau le plus général de la hiérarchie se trouvent des concepts très généraux comme "Anatomie" ou "Troubles Mentaux". Aux niveaux les plus bas se trouvent des concepts tels que "cheville", "trouble de conduite". Plus de 177 000 mots sont associés aux concepts permettant par exemple de retrouver pour le terme "Vitamine C" le concept "acide ascorbique".

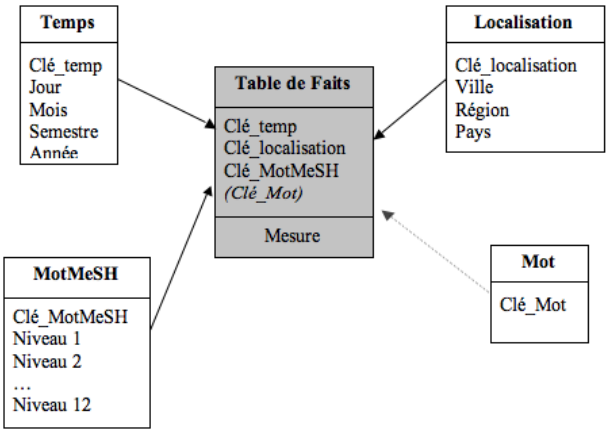

FIG. 3 – *Le schéma associé à l'application MeSH*

Le modèle de données est modifié afin de prendre en compte cette nouvelle dimension comme l'illustre la figure 3. Par rapport au modèle précédent (cf. figure 2) la dimension "Mot" a été remplacée par MotMeSH. MotMeSH est muni d'un ordre partiel, $\sqsubseteq_{motMeSH}$, pour pouvoir comparer les différentes valeurs du domaine.

Nous pouvons remarquer que ce modèle peut facilement être étendu (C.f. flèche en pointillé entre la dimension Mot et la table de faits) pour pouvoir prendre en compte les mots appris autour de ceux du MeSH de la même manière que dans la section 2.

L'un des problèmes principaux avec l'utilisation de ce thesaurus est que différents mots peuvent apparaître à plusieurs niveaux de la hiérarchie. Cette ambiguité pose le problème de l'utilisation des opérateurs de type Roll-up ou Drill-down pour naviguer dans le cube. De manière à illustrer notre problématique considérons l'exemple suivant.

Exemple 3 *Considérons le tweet suivant : "pneumonia & serious nerve problems. can't stand up. possible myasthenia gravis treatable with meds.". Si nous recherchons dans MeSH le concept associé à pneumonia, nous constatons que ce mot intervient à plusieurs endroits : dans la hiérarchie pneumonia $\leq$ respiratory tract infections $\leq$ respiratory tract diseases $\leq$ diseases ou dans la hiérarchie pneumonia $\leq$ lung diseases $\leq$ respiratory tract diseases $\leq$ diseases (C.f. Figure 4). En fonction de la position une opération de Roll-up sur pneumonia ne donnera pas le même résultat (i.e. "respiratory tract diseases" vs. "lung diseases").*

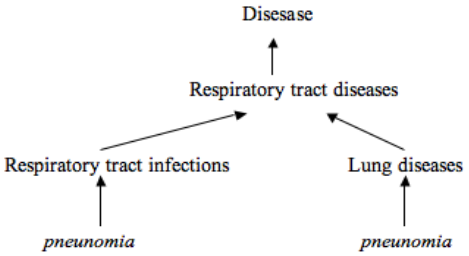

FIG. 4 – *Un exemple du thesaurus MeSH*

3.2 Comment repérer le contexte d'un tweet ?

Nous avons illustré, dans l'exemple 3, la difficulté de repérer le contexte d'un terme dans la hiérarchie. Cependant, si nous examinons le tweet associé, nous pouvons constater que les mots utilisés peuvent aider à déterminer le contexte. Il est ainsi évident que le sujet du tweet concerne davantage le concept "respiratory tract infections" que le concept "lung diseases". Aussi, par la suite, nous considérons l'hypothèse suivante : *plus le mot d'un tweet apparaît fréquemment avec le parent d'un concept dans le même contexte du tweet, plus le mot appartient à ce concept*.

De manière à désambiguïser les mots polysémiques de la hiérarchie de MeSH, nous avons adapté la méthode *AcroDef*$_{IM3}$ décrite dans (Roche et Prince (2008)) où les auteurs montre notamment l'efficacité de cette mesure dans un contexte biomédical de désambiguisation d'acronymes. Cette mesure qui est fondée sur l'Information Mutuelle au cube (Daille (1994)) calcule la dépendance de deux mots dans un contexte donné. Contrairement à l'Information Mutuelle, l'Information Mutuelle au cube privilégie les co-occurrences fréquentes. Par ailleurs, *AcroDef*$_{IM3}$ prend en considération le contexte dans lequel les co-occurrences sont présentes. En appliquant un contexte C (mots illustrant un contexte), l'approche *AcroDef*$_{IM3}$ est donnée par la formule ci-dessous.

$$AcroDef_{IM3}(m1, m2) = \frac{(nb(m1\ and\ m2\ and\ C))^3}{nb(m1\ and\ C) \times nb(m2\ and\ C)} \quad (2)$$

Dans notre cas, nous souhaitons calculer la dépendance entre un mot *m* à désambiguïser et différents mots m_t provenant des tweets, en utilisant le contexte de la hiérarchie (c'est-à-dire les parents *p* du mot m). Notre approche consiste à calculer la dépendance entre m et les autres mots m_t présents dans une fenêtre de cinq mots d'un tweet. Concrètement, nous considérons les deux mots précédant et suivant le mot m. Notons que nous ne prenons en compte dans ce processus que les noms, adjectifs et verbes qui ont été préalablement sélectionnés avec un étiqueteur grammatical. Ainsi, quatre requêtes du type '$m\ m_t\ p$' sont nécessaires, où p désigne les mots du contexte C des tweets. Alors, le nombre de résultats retournés par les quatre requêtes, noté $nb(m, m_t, p)$, nous permet de calculer la dépendance entre les mots d'un tweet, tout en prenant en compte le contexte de la hiérarchie tel que présenté dans la formule 2.

Exemple 4 *Considérons le mot 'pneumonia' à désambiguiser dans le tweet de l'exemple 3. Ici, nous calculons la dépendance entre ce mot m et les autres mots suivant 'pneumonia' : 'serious' et 'nerve'. Cette dépendance est calculée par rapport au contexte des deux pères possibles de la hiérarchie MeSH. Afin de prédire à quel élément du MeSH nous devons associer le mot 'pneumonia', nous effectuons les opérations suivantes :*

- $nb(pneumonia, m_t, \text{“}lung\ diseases'') = 227$ *(nombre de pages retourné avec les requêtes 'pneumonia serious "lung diseases"' et 'pneumonia nerve "lung diseases"')*
- $nb(pneumonia, m_t, \text{“}respiratory\ tract\ infections'') = 496$

La dépendance des mots est donnée par les formules ci-desous :

$$AcroDef_{MI3}^{\text{“}lung\ diseases''}(pneumonia, m_t)$$

$$= \frac{(nb(pneumonia, m_t, \text{“}lung\ diseases''))^3}{nb(pneumonia, \text{“}lung\ diseases'') \times nb(m_t, \text{“}lung\ diseases'')}$$

$$= 0.02$$

$$AcroDef_{MI3}^{\text{“}respiratory\ tract\ infections''}(pneumonia, m_t)$$

$$= \frac{(nb(pneumonia, m_t, \text{“}respiratory\ tract\ infections''))^3}{nb(pneumonia, \text{“}respiratory\ tract\ infections'') \times nb(m_t, \text{“}respiratory\ tract\ infections'')}$$

$$= 0.11$$

Ainsi, dans le tweet de l'exemple 3, pour le mot 'pneumonia', nous préférerons effectuer l'agrégation à partir du niveau du concept 'respiratory tract infections' du MeSH.

Notons que cette étape de désambiguisation qui se révèle indispensable pour les données de MeSH, est assez coûteuse en terme de nombre de requêtes à mener. Il semble donc plus pertinent d'appeler ces fonctions lors de la constitution de l'entrepôt plutôt qu'effectuer un tel traitement au moment de la navigation dans le cube.

4 Analyses de la mesure sur les tweets et premiers résultats

De manière à évaluer notre approche, différentes expérimentations ont été menées. Ces dernières ont été réalisées en utilisant Postgresql 8.4 via l'environnement Pentaho Mondrian 3.20.

Afin d'extraire les tweets utilisant les mots de MeSH, nous nous focalisons sur les messages associés à "virus diseases" (qui appartient à la hiérarchie "Disease"). Nous interrogeons Twitter à l'aide de tous les mots de la hiérarchie associée. Dans ce cadre, nous avons recueilli 1 801 310 tweets en anglais obtenus de janvier 2010 à février 2011. La répartition du nombre de tweets est présentée Figure 5.

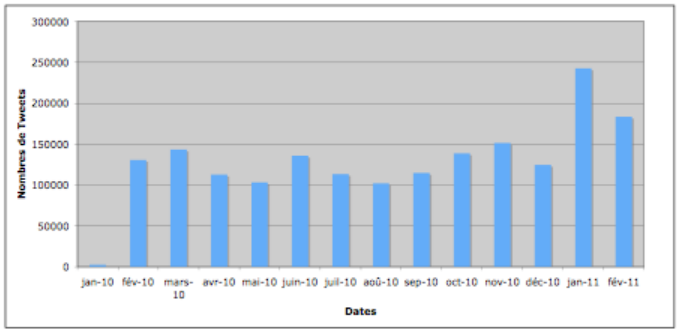

FIG. 5 – *Répartition du nombre de tweets de janvier 2010 à février 2011*

Les tweets récupérés sont pré-traités de la manière suivante. Tout d'abord, via l'API Twitter Stream [8], nous retenons la description *time zone* ou *user location* et complétons, via une base de localisations issues de Geoname [9], les informations relatives à la ville, à la région et au pays. Lorsque l'information est manquante dans la base, nous associons le tweet à la plus grande ville la plus proche. Les différents tags (i.e. RT, #, @) et liens sont ensuite extraits et la langue déterminée via TextCat [10]. Enfin, nous appliquons l'étiqueteur grammatical TreeTagger [11] pour extraire le type grammatical et le lemme associé à chaque élément du tweet. La visualisation des résultats est faite via le service Google Public Data Explorer de Google.

Etats Unis	**Illinois**	**Chicago**
wart	risk	risk
pneumonia	vaccination	wart
vaccination	wart	pneumonia
risk	pneumonia	wood
lymphoma	wood	colonoscopy
common cold	colonoscopy	x-ray
disease	x-ray	death
meningitis	encephalitis	school
infection	death	vaccination
vaccine	school	eye infection
life	eye infection	patient
hepatitis	man	russia

TAB. 1 – *Les 12 mots les plus fréquents en fonction du TF-IDF adaptatif au cours du mois de janvier 2011.*

Dans un premier temps, nous étudions l'impact de notre mesure sur les résultats puis nous

8. http ://dev.twitter.com/pages/user_streams
9. http ://www.geonames.org/
10. http ://www.let.rug.nl/˜vannoord/TextCat/
11. http ://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

montrons un exemple d'application.

Ainsi, dans ces expérimentations, nous analysons les premiers mots retournés par le TF-IDF adaptatif (scores les plus élevés). Par exemple, le tableau 1 présente les 12 premiers mots (top-12) des tweets des Etats-Unis, pour l'Etat de l'Illinois et pour la ville de Chicago au cours du mois de janvier 2011. Le tableau 2 présente quant à lui les mots de tous les pays sur toute la période.

pneumonia	wart	hepatitis
die	help	rabies
month	son	leukemia
age	smallpox	awareness
hope	meningitis	influenza
disease	cure	virus
common cold	treatment	sick

TAB. 2 – *Les 21 mots les plus fréquents en fonction du TF-IDF sur la période d'analyse*

Dans ce cadre, nous souhaitons identifier si les n mots retournés par le TF-IDF adaptatif sont les mêmes que ceux retournés par le TF-IDF classique. Le tableau 3 montre le pourcentage de mots exclusivement retournés par le TF-IDF adaptatif sur la base des niveaux "Pays" et "Villes" (moyenne des 30 villes les plus importantes en terme de nombre de mots). Par exemple, ce tableau met en relief que 12,22% des mots du top-3 obtenus par le TF-IDF adaptatif ne sont pas présents dans le top-3 des mots retournés par le TF-IDF classique. Le nombre de mots retournés par notre mesure adaptative est légèrement plus important si nous considérons les dix premiers mots (top-1, top-3, top-5, top-10). Pour les pays, nous pouvons remarquer que les premiers mots retournés diffèrent globalement moins avec notre mesure.

top-n	**Top 30 des villes**	**Tous les pays**
top-1	16,67%	0,00%
top-3	12,22%	0,00%
top-5	13,33%	0,00%
top-10	10,33%	10,00%
top-50	6,27%	10,00%
top-100	5,73%	4,00%
top-150	5,42%	6,67%
top-200	4,98%	5,50%
top-250	4,79%	4,40%
top-300	5,06%	5,67%
top-350	5,24%	7,14%
top-500	5,21%	6,00%

TAB. 3 – *Pourcentage de mots exclusivement retournés par le TF-IDF adaptatif.*

Nous pouvons noter que les mots retournés sont majoritairement identiques en utilisant les deux mesures. L'utilisateur peut alors accorder une confiance de pertinence élevée à de tels mots. Ceci est par exemple le cas pour les premiers mots retournés pour les villes de Londres (*wart*) et Boston (*pneumonia*). A contrario, les mots plus spécifiquement retournés

par la mesure adaptative (par exemple, le mot *leukemia* propre à Atlanta) mettent davantage en relief la discriminance par rapport à l'ensemble des villes.

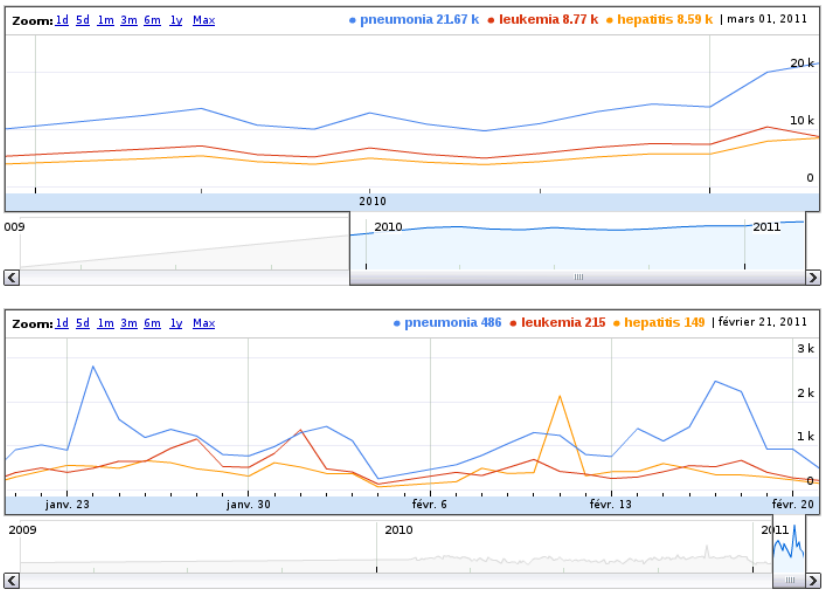

FIG. 6 – *Répartition de l'utilisation des mots pneumonia, leukemia et hepatitis sur la période considérée (haut) et détail sur les mois de janvier et février (bas)*

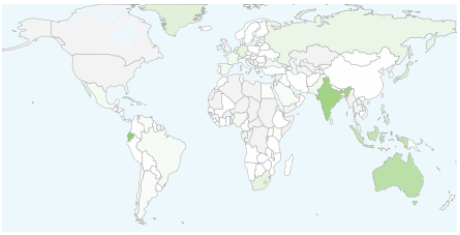

FIG. 7 – *Répartition de l'utilisation du mot hepatitis*

FIG. 8 – *Répartition de l'utilisation du mot leukomia*

Nous considérons à présent un exemple d'application de notre approche. La figure 6 présente la répartition des mots *hepatitis*, *leukomia* et *pneumonia* au cours de la période considérée (hors Etats Unis [12]). Les figures *7*, *8* et *9* visualisent la couverture mondiale de ces mots excluant les États-Unis, le Royaume-Uni et le Canada. Cette couverture est obtenue en fixant la dimension localisation et en examinant la fréquence du mot sur la période considérée. La

12. Les expériences menées montrent qu'actuellement la plupart des tweets sont émis des Etats Unis.

FIG. 9 – *Répartition de l'utilisation du mot pneumonia*

figure 10 montre un exemple de répartition des tweets en se focalisant sur l'utilisation du mot *pneumania* au Royaume-Uni.

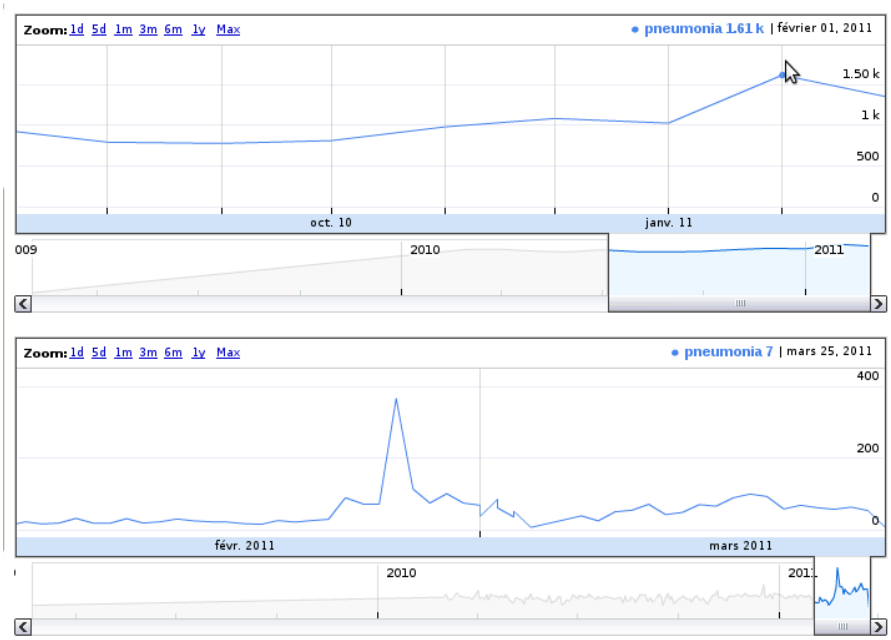

FIG. 10 – *Répartition de l'utilisation du mot pneumonia dans les tweets au Royaume-Uni sur la période fin d'année 2010-début 2011*

5 Etat de l'art

L'analyse des données textuelles issues des tweets est un domaine de recherche actuel et de nombreuses propositions existent. Par exemple, dans Sakaki et al. (2010), les auteurs

proposent d'analyser le contenu des tweets pour détecter en temps réel des alarmes lors d'apparitions de tremblement de terre. Les auteurs de TwitterMonitor (Mathioudakis et Koudas (2010)) présentent un système pour extraire automatiquement les tendances dans le flot des streams. Une approche assez similaire est proposée dans Benhardus (2010). Cependant, à notre connaissance, la plupart des travaux existants proposent un traitement particulier des tweets et n'offrent pas d'outils généraux permettant au décideur, en fonction de ses besoins, de pouvoir manipuler l'information contenue dans les tweets. Ainsi, il n'existe que peu de travaux qui se soient intéressés à l'utilisation de cubes pour les tweets. Quelques travaux récents se sont par contre intéressés à intégrer les données textuelles dans un contexte d'entrepôt de données. Dans ce cadre, des méthodes d'agrégation adaptées aux données textuelles ont été proposées. Par exemple, les travaux de Keith et al. (2005) proposent d'utiliser des approches de TALN (Traitement Automatique du Langage Naturel) pour agréger les mots ayant la même racine ou les mêmes lemmes (connaissances morpho-syntaxiques). Les auteurs proposent également de rassembler les mots sur la base de classifications sémantiques généralistes existantes (WordNet et Roget). Outre l'utilisation de connaissances morpho-syntaxiques et sémantiques, d'autres travaux utilisent des approches numériques issues du domaine de la Recherche d'Information (RI) pour agréger les données textuelles (Pujolle et al. (2008); Lin et al. (2008); Pérez-Martínez et al. (2008)). Ainsi, Lin et al. (2008) agrègent les documents sur la base des mots-clés présents dans ces derniers en utilisant une hiérarchie sémantique des mots présents dans l'entrepôt et des mesures issues de la RI. De telles méthodes issues de la RI sont aussi utilisées dans les travaux de Pérez-Martínez et al. (2008) qui consistent à prendre en compte une dimension "contexte" et "pertinence" pour construire un entrepôt de données textuelles appelé R-Cube. Certaines approches proposent d'ajouter une nouvelle dimension spécifique. Par exemple, dans Zhang et al. (2009), les auteurs ajoutent une dimension 'topic' et appliquent l'approche PLSA (Hofmann (1999)) pour extraire les thèmes représentatifs des documents dans cette nouvelle dimension. Enfin, Pujolle et al. (2008) proposent d'agréger des parties de documents afin d'offrir au décideur des mots caractéristiques propres à cette agrégation. Dans ce cadre, les auteurs utilisent une première fonction pour sélectionner les mots les plus significatifs en utilisant la mesure $TF\text{-}IDF$ classique issue du domaine de la RI. Une approche assez similaire est proposée dans Benhardus (2010). L'objectif de nos travaux est assez similaire à ces dernières approches. Toutefois, dans notre cas, nous souhaitons proposer une mesure qui tienne compte d'une hiérarchie existante (i.e. localisation) dans les mots pertinents restitués aux décideurs. En d'autres termes, nous souhaitons ne retourner que les mots significatifs (i.e. top-k) par rapport à un niveau donné. Par exemple, dans le cas des mots significatifs, leur agrégation *Avg-Kw* qui utilise un $TF\text{-}IDF$ ne permet de connaître que les mots significatifs à un niveau bas de la hiérarchie mais ne permet pas de prendre en compte une hiérarchie existante.

De manière analogue aux travaux sur les messages issus de twitter proposés par Cheong et Lee (2009), notre système permet de détecter des tendances liées aux thèmes, lieux géographiques, etc. Notons que notre approche qui intègre les messages dans des entrepôts de données prend en compte des informations hiérarchiques propres aux tweets. Dans un tel contexte, le décideur peut mener une analyse plus fine des données.

6 Conclusion

Dans cet article nous avons proposé une nouvelle approche pour analyser les tweets à partir de leurs caractéristiques multidimensionnelles. L'originalité de notre proposition est de pouvoir définir et manipuler des cubes de tweets. Nous avons ainsi montré au travers de deux modèles et applications différentes : sans hiérarchie prédéfinie sur les tweets (i.e. analyse des régimes) et avec hiérarchie existante (i.e. en utilisant le thesaurus MeSH), que l'analyse des tweets nécessitait la définition de nouvelles mesures et qu'une étape de contextualisation était pertinente. Les expérimentations menées ont souligné l'émergence de points nouveaux et intéressants pour le décideur qui n'auraient pas pu être découverts dans un autre cadre.

Les perspectives associées à ce travail sont nombreuses. Tout d'abord nous souhaitons étendre l'approche proposée à la prise en compte des opinions exprimées dans les tweets. De récents travaux analysent l'humeur des personnes (e.g. http ://twittermood.org/). Nous souhaitons enrichir ces approches en analysant le contenu des tweets et ainsi pouvoir automatiquement extraire des connaissances du type : quelles sont les personnes ayant suivies un régime et qui n'en sont pas satisfaites ? Dans un second temps, nous souhaitons considérer les tweets comme étant disponibles sous la forme d'un flot.

Références

Benhardus, J. (2010). Streaming trend detection in twitter. In *National Science Foundation REU for Artificial Intelligence, Natural Language Processing and Information Retrieval*, University of Colarado.

Bringay, S., A. Laurent, P. Poncelet, M. Roche, et M. Teisseire (2010). Bien cube, les données textuelles peuvent s'agréger ! In *Actes de la conférence 'Extraction et gestion des connaissances' (EGC)*, pp. 585–596.

Buckley, C., A. Singhal, et M. Mitra (1995). New retrieval approaches using smart : Trec 4. In *TREC*.

Cheong, M. et V. Lee (2009). Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *Proceeding of the 2nd ACM workshop on Social web search and mining*, SWSM '09, pp. 1–8.

Codd, E., S. Codd, et C. Salley (1993). Providing OLAP (on-line analytical processing) to user-analysts : An IT mandate. In *White Paper*.

Daille, B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filltres linguistiques. Technical report, Phd Thesis, University Paris VII, France.

Grabs, T. et H.-J. Schek (2002). Eth zürich at inex : Flexible information retrieval from xml with powerdb-xml. In *XML with PowerDB-XML. INEX Workshop*, pp. 141–148. ERCIM Publications.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *In Proc. of Uncertainty in Artificial Intelligence, UAI'99*, pp. 289–296.

Jin, R., C. Falusos, et A. G. Hauptmann (2001). Meta-scoring : automatically evaluating term weighting schemes in ir without precision-recall. In *SIGIR '01 : Proceedings of the 24th*

annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, pp. 83–89. ACM.

Keith, S., O. Kaser, et D. Lemire (2005). Analyzing large collections of electronic text using OLAP. Technical Report TR-05-001, UNBSJ CSAS.

Lin, C. X., B. Ding, J. Han, F. Zhu, et B. Zhao (2008). Text Cube : Computing IR Measures for Multidimensional Text Database Analysis. In *In Proc. of Int. Conf. on Data Mining (ICDM'08)*, pp. 905–910.

Mathioudakis, M. et N. Koudas (2010). Twittermonitor : trend detection over the twitter stream. In *Proceedings of 2010 International Conference on Management of Data (SIGMOD 2010), DŐmostration*.

Pérez-Martínez, J. M., R. B. Llavori, M. J. A. Cabo, et T. B. Pedersen (2008). Contextualizing data warehouses with documents. *Decision Support Systems 45*(1), 77–94.

Pujolle, G., F. Ravat, O. Teste, et R. Tournier (2008). Fonctions d'agrégation pour l'analyse en ligne (OLAP) de données textuelles. fonctions top_kwk et avg_kw opérant sur des termes. *Ingénierie des Systèmes d'Information 13*(6), 61–84.

Robertson, S. E. et S. Walker (1999). Okapi/keenbow at trec-8. In *TREC*.

Roche, M. et V. Prince (2008). Managing the acronym/expansion identification process for text-mining applications. *International Journal of Software and Informatics, Special issue on Data Mining 2*(2), 163–179.

Sakaki, T., M. Okazaki, et Y. Matsuo (2010). Earthquake Shakes Twitter Users : Real-time Event Detection by Social Sensors. In *Proceedings of 19th World Wide Web Conference (WWW 2010)*.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Commun. ACM 18*(11), 613–620.

Zhang, D., C. Zhai, et J. Han (2009). Topic cube : Topic modeling for OLAP on multidimensional text databases. In *In Proc. of the SIAM Int. Conference on Data Mining*, pp. 1123–1134.

Summary

Exchanged tweets on the Internet are an important information source, even if their characteristics make them difficult to analyze (a maximum of 140 characters, shorthand notations, ...). In this paper, we define a model of data warehouse to develop and analyze large volumes of tweets by proposing relevant measures in a knowledge discovery context. Using data warehouses in order to store and analyze textual documents is not new. Traditionally they adapt classical measures which are not really adapted to the data specificities. Furthermore we propose that, if a hierarchy is available, we can automatically detect the context. Conducted experiments on real data show the relevance of our approach.