Analyse OLAP sur des tweets et des blogs : un retour d'expérience

Brice Olivier, Cécile Favre, Sabine Loudcher

Laboratoire ERIC, Université de Lyon, 5 Avenue Pierre Mendès-France
69676 Bron Cedex, FRANCE
{brice.olivier, cecile.favre, sabine.loudcher}@univ-lyon2.fr

Résumé. Le projet ANR IMAGIWEB dans lequel s'inscrit ce travail s'est donné pour mission d'étudier les images véhiculées sur Internet en se basant sur la détection d'opinions. Deux cas d'étude ont été définis : (1) le premier vise à répondre aux besoins d'analyse de chercheurs en science politique grâce à des données issues de Twitter durant la campagne présidentielle de 2012 ; (2) le second doit permettre à l'entreprise française EDF d'évaluer l'opinion du public en matière de sécurité, d'emploi et de prix à partir de billets de blogs. Dans cet article, nous présentons un retour d'expérience sur l'usage de l'analyse en ligne OLAP (OnLine Analytical Processing) pour des données textuelles, mettant en avant l'intérêt de ce type d'analyse pour les membres du projet.

1 Introduction

Le projet ANR IMAGIWEB [1] consiste à analyser et à suivre l'évolution de l'image (au sens de l'opinion) sur la toile, d'une part des personnages politiques à travers le réseau social Twitter, et d'autre part de l'entreprise EDF vis-à-vis du nucléaire en utilisant des blogs comme données. Ce projet regroupe différents partenaires parmi lesquels un laboratoire de recherche en science politique, des entreprises et des laboratoires de recherche en fouille de données.

Dans un premier temps, les tweets et blogs récoltés sont annotés manuellement pour relater l'opinion qu'ils véhiculent. Par la suite, l'enjeu sera de détecter automatiquement les opinions grâce à des méthodes de fouille d'opinion. Au-delà de la détection des opinions, pour mieux comprendre et analyser le contenu des tweets et des blogs, l'enjeu est aussi de les visualiser et de les explorer. Ainsi, un autre objectif du projet consiste à fournir à l'utilisateur, qu'il soit politologue, sociologue, marketeur ou encore analyste, un outil pour explorer les données (issues de tweets ou de blogs) et pour analyser en ligne l'opinion selon différents points de vue (sujets, temps, ...). L'analyse OLAP (OnLine Analytical Processing) permet de répondre à cet objectif de navigation, d'analyse et de visualisation.

L'OLAP sur des données textuelles correspond à une thématique de recherche récente avec des enjeux scientifiques importants. En effet, si l'OLAP a su montrer tout son potentiel analytique sur des "données classiques", la prise en compte de données textuelles nécessite une adaptation ou une évolution de l'OLAP pour prendre en compte les spécificités de ces données

1. ANR 2012-CORD- 002-01

(Ravat et al., 2007; Zhang et al., 2009). Quelques travaux de recherche encore plus récents portent sur l'analyse OLAP de tweets, un cas particulier de données textuelles (Ben Kraiem et al., 2014; Bringay et al., 2011). Dans ce contexte, l'objectif de ce papier est de (1) démontrer l'intérêt de l'analyse OLAP pour ce type de données en se basant sur des cas d'étude réels, (2) relater une implémentation concrète "classique" en utilisant des outils existants.

Pour ce faire, dans la section 2 nous commençons par présenter les deux cas d'étude. Dans la section 3, nous évoquons les aspects de modélisation multidimensionnelle et de navigation. Dans la section 4, nous exposons la mise en œuvre, avant de conclure dans la section 5.

2 Deux cas d'étude

Dans le cadre du projet IMAGIWEB, deux cas d'étude sont traités : des tweets à caractère politique et des billets de blogs traitant de l'entreprise EDF et du nucléaire. Pour chacun des cas, un processus d'annotation manuelle concernant l'opinion véhiculée a été mis en place.

Données tweets et besoins d'analyse Dans le cadre du projet IMAGIWEB, les tweets ont été recueillis grâce à l'API *Streaming* de Twitter. Ce sont des tweets en français, à caractère politique, portant sur Nicolas Sarkozy et François Hollande, avant et après les élections présidentielles de 2012. Les données extraites sont le contenu du tweet, le pseudonyme du twittos, la date du tweet, l'image (à savoir François Hollande ou Nicolas Sarkozy, c'est à dire l'entité sur laquelle porte le tweet), l'URL qui mène vers le tweet.

Une annotation est faite par un annotateur sur un extrait ou un passage d'un tweet. L'annotateur détermine l'opinion contenue dans le passage (avec une polarité allant de -2 pour une opinion très négative à +2 pour une opinion très positive en passant par le zéro si l'opinion est neutre ou par le NULL s'il n'y a pas d'opinion) ainsi que la cible (le sujet sur lequel porte le passage) et la sous-cible. Les cibles et sous-cibles ont été déterminées par les membres du projet. Citons comme exemple de cible "bilan", "compétences", "positionnement". Pour la cible "positionnement", les sous-cibles sont "alliance", "écologie", "économie" et "sociétal". Enfin l'annotateur donne un niveau de confiance dans son annotation. 4073 tweets ont été annotés manuellement, ce qui a donné lieu à 5674 annotations.

Les données tweets constituent le terrain d'analyse des chercheurs en science politique et en sociologie. Les politologues souhaitent pouvoir suivre l'évolution dans le temps des deux images que sont François Hollande et Nicolas Sarkozy à travers Twitter. L'analyse de ces données, à la fois des tweets eux-mêmes et de leurs annotations, constitue un premier enjeu du projet.

Données blogs et besoins d'analyse Les blogs à analyser concernent tout ce qui touche à EDF et au nucléaire. À partir d'un ensemble de blogs, tous les articles, en français, avec au moins une occurrence du sigle EDF ou des mots "Electricité de France" et de "nucléaire" ont été collectés. Les données contiennent le titre de l'article, l'URL du site web dont provient l'article, la date, le contenu textuel et l'image (sécurité, emploi ou prix). Les données blogs contiennent également le passage annoté (à chaque article correspond un ou plusieurs passages), la cible ("politique", "tarifs" ou encore "risques"), la sous-cible (par exemple "démantèlement/durée de vie" ou "expertise/incident" pour la cible "risques"), la polarité et la confiance. 560 articles ont été annotés manuellement en 3420 annotations (6,1 annotations par article en moyenne).

Par rapport aux besoins, les marketeurs d'EDF souhaitent centrer leur analyse sur les notions de cibles, de polarité. Ils souhaitent également pouvoir naviguer dans les données selon le type de structure (organisme) dont est issu le blog. Cette information peut être portée par l'extension du site web. Par exemple, une organisation à but non lucratif aura généralement un site web avec l'extension ".org" alors qu'une société aura un site web avec une extension ".com".

3 Modélisation et analyse OLAP

Partie intégrante des systèmes d'information décisionnels, l'OLAP repose sur un modèle multidimensionnel avec les concepts de faits, mesures, et dimensions. Les faits sont les objets que l'on cherche à analyser. Ils sont décrits ou évalués par des indicateurs ou mesures et sont observables selon plusieurs dimensions ou axes d'analyse. Les axes d'analyse peuvent comporter plusieurs niveaux de granularité de l'information, organisés en une ou plusieurs hiérarchies de dimension.

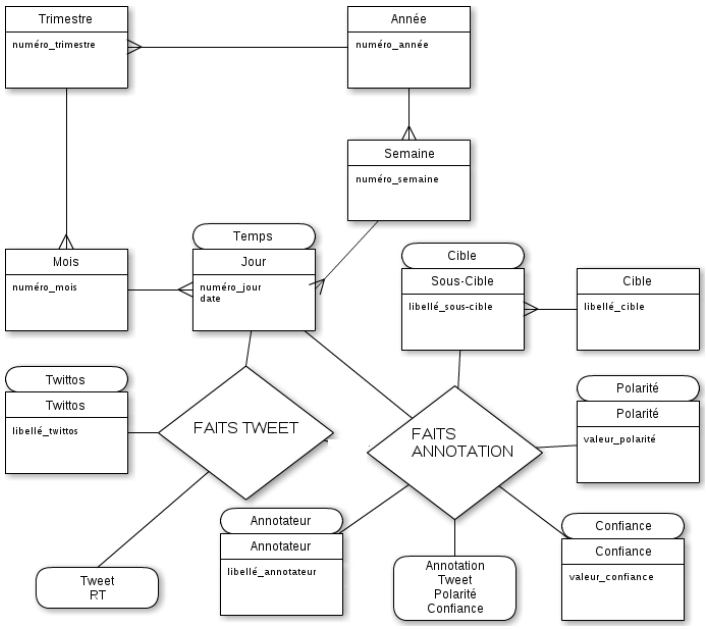

FIG. 1 – *Modèle multidimensionnel des tweets, notation de Malinowski et Zimányi (2007)*

Dans le modèle associé aux tweets (cf. figure 1), deux faits sont observés : *ANNOTATION* et *TWEET*. À ces faits sont associées plusieurs dimensions, comme le *temps*, la *cible* ou encore l'*annotateur*. Dans la dimension *temps*, on retrouve plusieurs niveaux de granularité de l'information avec deux hiérarchies : *{jour, semaine, année}* et *{jour, mois, trimestre, année}*.

Le fait *TWEET* va permettre de compter le nombre de tweets et de *RT*[2] en fonction de

2. La mention *RT* au début d'un tweet signifie que le tweet provient d'une autre personne et qu'il a été *retweeté*

leur auteur et du temps à différents niveaux de détails. Le fait *ANNOTATION*, associé aux annotations sur les tweets, peut être évalué grâce à différentes mesures : *Annotation* permettra d'obtenir le nombre d'annotations ; *Tweet* se focalise sur le nombre de tweets réels ; *Polarité* (qui sera exploitée grâce aux fonction d'agrégat moyenne et somme) traduit l'opinion ; *Confiance* donne un indice quant à la confiance accordée à la polarité par l'annotateur. Le modèle permet également de retrouver l'*image*, la *cible*, l'*annotateur*, et bien sûr, le *temps*. La dimension *annotateur* rendra possible la comparaison de l'annotation automatique à celle manuelle le moment venu. Les dimensions *image*, *cible* et *temps* sont cruciales pour l'analyse. Une des particularités de ce modèle est de retrouver la polarité et la confiance aussi bien en mesure qu'en dimension. Cela permet de visualiser les données selon différentes manières. La polarité en tant que dimension permet par exemple de visualiser le nombre de fois où la polarité +2 est affectée alors qu'en la plaçant en tant que mesure, elle peut être agrégée avec des fonctions comme la somme ou la moyenne.

Le modèle pour les blogs est assez similaire à celui des tweets. On retrouve deux faits *Article* et *Annotation*. Les mesures et les fonctions d'agrégat associées sont identiques. On retrouve également plusieurs dimensions en commun, à savoir la polarité, la confiance, la cible et le temps. Toutes ces notions similaires sont en fait celles associées au besoin commun concernant l'analyse de l'opinion. En revanche, notons comme différence que les blogs disposent d'un titre grâce à la dimension *Blog* et qu'ils sont également porteurs d'informations sur la structure qui héberge l'article (grâce à l'extension du site web) via une dimension *Structure*.

À partir du modèle multidimensionnel, pour introduire la navigation, la notion de cube OLAP est utilisée. Ainsi, deux cubes ont été créés concernant les données issues de Twitter et il en est de même pour les blogs. La navigation se caractérise par l'application d'opérateurs tels que le *Drill Down* qui permet d'aller vers un niveau plus détaillé selon la hiérarchie de dimension définie préalablement dans le modèle, en appliquant une fonction d'agrégat sur la mesure qui est observée. Il s'agit par exemple de passer de l'observation de la polarité moyenne par trimestre à l'observation par mois selon la hiérarchie temporelle. L'opérateur inverse s'appelle le *Roll Up*. Notons également l'existence de l'opérateur *Slice & Dice* qui permet de sélectionner certaines valeurs pour certains axes d'analyse. Par exemple, dans un cube qui permet d'observer le nombre de tweets par mois et par cible, il serait possible de sélectionner quelques cibles sur lesquelles on souhaite se focaliser.

4 Mise en œuvre

Dans le cadre du projet IMAGIWEB, nous avons retenu MySQL comme SGBD en raison de contraintes techniques du projet. Nous avons également choisi de développer notre propre ETL (*Extract Transform Load*, phase correspondant à l'alimentation des données) car nous souhaitions pouvoir apporter des transformations très particulières en lien avec le contenu textuel (relatives à la fouille de texte) pour la suite du projet. Enfin, nous avons préféré le serveur OLAP Pentaho Mondrian en lui greffant l'interface graphique Saiku pour l'étendue de sa communauté et la prise en main de son environnement. L'implémentation résultante permet de naviguer dans les données en construisant des tableaux de bord très facilement pour l'utilisateur comme nous l'illustrons par la suite sur les données Twitter. Notons qu'il y a un menu sur l'interface qui permet également de représenter les données issues de la navigation sous forme de différents types de graphiques qui sont générés très simplement par l'utilisateur.

Initialement, le politologue peut par exemple observer la polarité moyenne en fonction du temps en trimestre pour les entités Hollande et Sarkozy. Puis, pour observer de façon plus précise, il peut obtenir le détail par mois (ce qui correspond au niveau OLAP à une opération de *Drill Down*), en se focalisant simplement sur Hollande (réalisant ainsi une opération de *Slice*), obtenant ainsi les résultats figurant dans la figure 2.

Ainsi, le politologue peut constater une baisse importante de popularité entre le mois de Mai et le mois de Juin (polarité de -0.346 à -0.658). Il peut ensuite détailler les cibles sur lesquelles cette baisse est plus importante, en ajoutant la dimension *Cible* dans les résultats. Le tableau 1 qui en résulte permet d'observer que l'opinion des Twittos a particulièrement diminué sur ses performances (-0.294 à -1.000) mais aussi sur son positionnement et son projet.

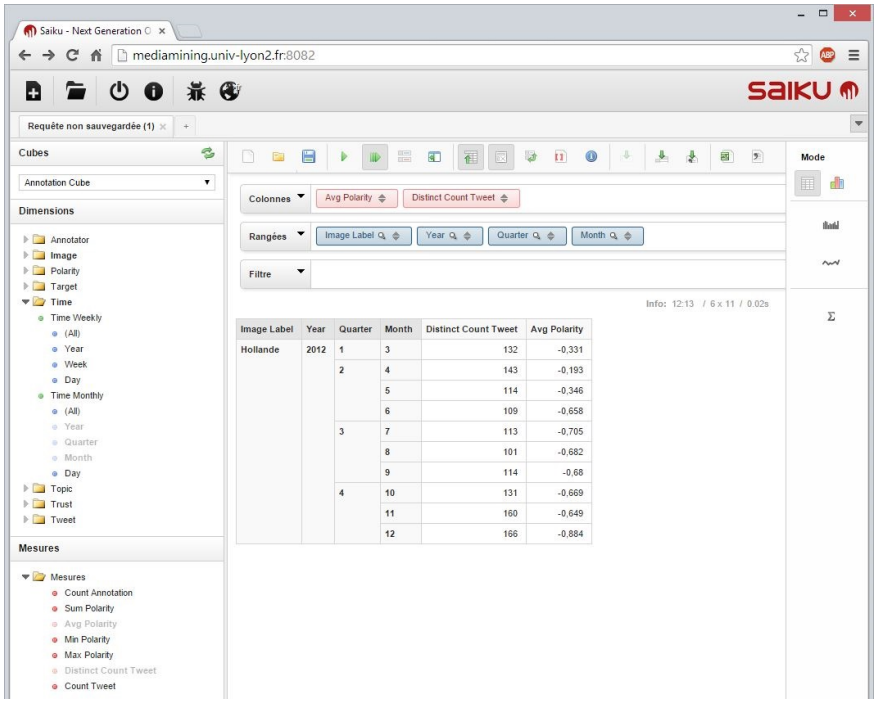

FIG. 2 – *Polarité moyenne et nombre de tweets en fonction du temps en mois pour Hollande*

L'intérêt pour le chercheur en science politique est ici, sur la base de la navigation, de pouvoir établir des liens entre l'opinion exprimée sur le web et des évènements de la vie politique, d'observer également à quel point le Web est un miroir ou non de l'opinion publique au sens large (comparaison avec les sondages d'opinion classiques).

5 Conclusion

Dans le cadre du projet IMAGIWEB, l'analyse OLAP était une des pistes à explorer pour visualiser les données. La mise en œuvre de l'architecture décisionnelle a répondu à de réels

Cible	Mois	Polarité moyenne	Nombre de tweets
Entité	5	-0.438	41
	6	-0.400	25
Performances	5	-0.294	18
	6	-1.000	28
Positionnement	5	-0.667	22
	6	-0.857	44
Projet	5	-0.200	22
	6	-0.737	21

TAB. 1: Extrait des résultats par cible pour Mai et Juin pour Hollande

besoins pour les politologues et marketeurs d'EDF de navigation dans des données textuelles. L'outil mis en œuvre est convivial et simple d'utilisation pour explorer des données volumineuses ne pouvant être analysées manuellement, données étant amenées à l'être davantage dans la suite du projet.

L'intérêt de cet article a été ainsi de fournir un retour d'expérience sur l'analyse OLAP de tweets et de blogs dans le cadre d'un projet concret avec de vraies attentes pour les partenaires. L'enjeu est alors à présent d'aller vers une exploitation accrue du contenu textuel dans ce travail d'analyse en ligne.

Références

Ben Kraiem, M., J. Feki, K. Khrouf, F. Ravat, et O. Teste (2014). OLAP of the tweets : From modeling to exploitation. In *IEEE RCIS, Marrakesh, Morocco*, pp. 45–55.

Bringay, S., N. Béchet, F. Bouillot, P. Poncelet, M. Roche, et M. Teisseire (2011). Analyse de gazouillis en ligne. In *EDA, Clermont-Ferrand, France*, Volume B-7 of *RNTI*, pp. 87–102.

Malinowski, E. et E. Zimányi (2007). Logical representation of a conceptual model for spatial data warehouses. *GeoInformatica 11*(4), 431–457.

Ravat, F., O. Teste, et R. Tournier (2007). OLAP Aggregation Function for Textual Data Warehouse. In *ICEIS, Funchal, Madeira - Portugal*, Volume DISI, pp. 151–156.

Zhang, D., C. Zhai, et J. Han (2009). Topic cube : Topic modeling for olap on multidimensional text databases. *SDM 9*, 1124–1135.

Summary

The IMAGIWEB ANR project is focused on studying the images portrayed on the Internet based on the detection of opinions. Two case studies were identified: (1) the first is designed to meet the analytical needs of researchers in policy science through data from Twitter during the 2012 presidential campaign ; (2) the second is to allow the French company EDF to be able to assess the public's opinion on safety, employment and prices from blog posts. Here we present a feedback on the use of OLAP (OnLine Analytical Processing) for textual data on these two case studies, demonstrating the interest for the project members.