



UNIVERSITÉ
LUMIÈRE
LYON 2
UNIVERSITÉ DE LYON

CommentWatcher

plateforme Web open-source pour analyser
les discussions sur des forums en ligne

Marian-Andrei RIZOIU

BLEND 2013

2^{ème} octobre 2013

Lyon, France

Contexte

Laboratoire ERIC

Université Lumière

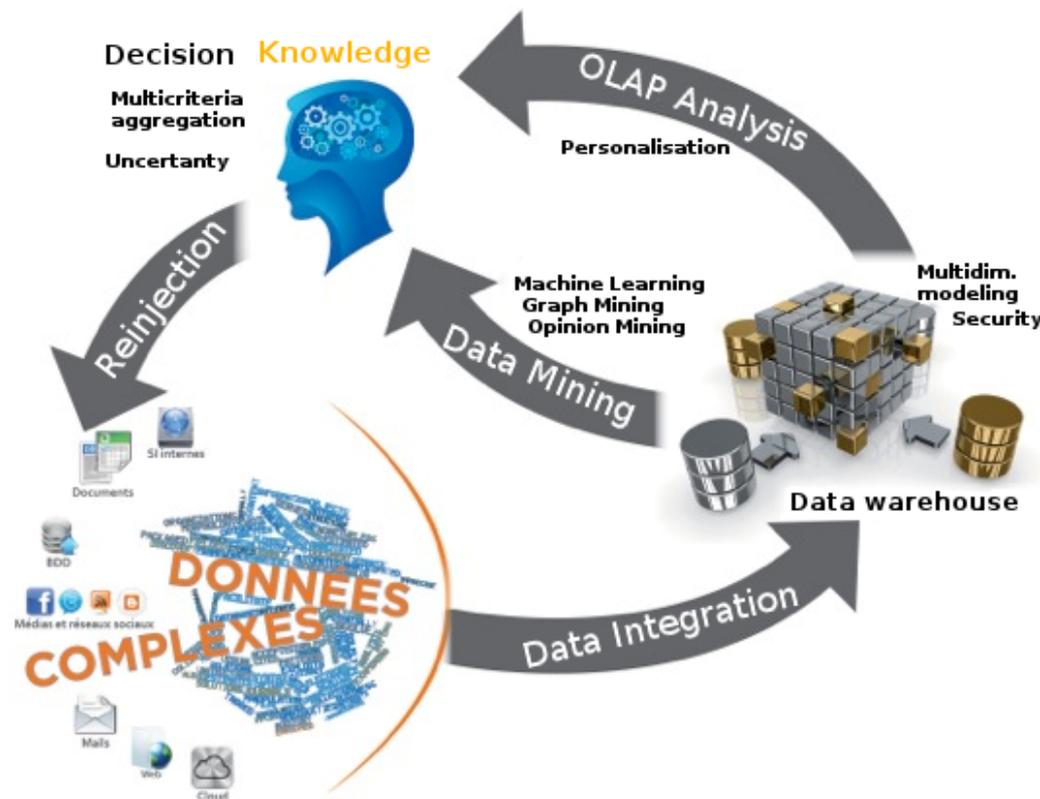
Sciences Humaines et Sociales

Sociologie, Psychologie, Linguistique, Histoire, etc.)



Projets de recherche multidisciplinaire

- Identifier des patrons à partir des textes mortuaires (*historiens*)
- Analyser des débats publiques : forums en ligne et média traditionnel (*sciences sociales*)
- Discours sur la médecine nucléaire: évolution diachronique et diastratique (*linguistes*)
- Evolution de l'images des politiciens et des entreprises à travers le média social (*sciences politiques*)
- Détecter des rôles sociaux dans des réseaux sociaux enferred à partir des discussions sur des forum en ligne (*Technicolor*)



Les données : Une collection de textes en langage naturel, souvent issus de l'internet.

Les défis :

- grands volumes de données ;
- besoin de résumer les « idées » principales : *les thématiques*
- besoin d'analyser comment les utilisateurs interagissent par rapport à l'information textuelle (diffusion de l'information, réseaux sociaux).

Nouvelle hausse du prix du tabac en juillet, jusqu'à 7 euros le paquet de cigarettes

Le HuffPost/AFP | Publication: 12/06/2013 09h09 CEST | Mis à jour: 12/06/2013 10h33 CEST

Like

Share 0



30 4 0
partager tweeter envoyer

SUIVRE: Smoking, Video, Marisol Touraine, Ac Prix, Santé, Santé, Tabac, Actualités

SANTÉ - Le prix des paquets de cigarets en juillet, a déclaré mercredi la ministre intervenue début juillet" et se fera "a p itélé.

L'hypothèse d'une hausse en deux te octobre- est donc abandonnée. Prévu sociale 2012 cette hausse ferait passer et celui des plus vendus à 7 euros.

SUPER UTILISATEUR DU HUFFPOST
dieu
295 Fans Suivre

il y a 19 minutes (11h09)
Pourquoi taxer un fumeur ? pourquoi ne pas détruire les plantations et éliminer les buralistes ?
Répondre Lien permanent | Partagez

LUMINET
17 Fans

il y a 12 minutes (11h16)
Pourquoi subventionner les producteurs de tabac???
Répondre Lien permanent | Partagez

cpamafaute
2 Fans

il y a 39 minutes (10h49)
Continuons d'appauvrir les Français par des taxes imbéciles qui ne changeront rien aux comportements des fumeurs ! Entre toutes les taxes et les impôts comment allons nous faire ? On ne cesse de nous parler de relance économique et on détruit le pouvoir d'achat des Français ! Vous pensez que lorsque toutes les entreprises seront fermées l'argent rentrera dans les caisses de l'état ????? Cette politique est un désastre pour notre pays.
Répondre Lien permanent | Partagez

Une solution issue du *data mining* :

- Extraire les thématiques des discussions
- Associer à ces thématiques des noms compréhensibles pour les humains
- Analyser le réseau social en ligne des utilisateurs qui discutent

Dimension appliquée :

- Demande très forte de la part de chercheurs, surtout dans le domaine des **Sciences Humaines et Sociales** (*Sociologie, Psychologie, Linguistique, Histoire, etc.*)
- Solution implémentée dans une plateforme d'analyse des discussion sur des forums en ligne **CommentWatcher**

Le contexte du travail - les forums de discussion

Difficultés :

- La plupart des outils ne traitent pas l'aspect réseau social des données forum [AME12, GUI13]
- Manque de jeux de données issues de forums
- Structure des sites qui change constamment
- Problème de licence sur le contenu des forums

The screenshot shows a forum thread with several posts. Red arrows and boxes highlight specific elements:

- Nom d'utilisateur:** Points to the name 'nicolas22' in the first post.
- Date du message:** Points to the timestamp 'il y a 54 minutes (11h47)' in the first post.
- Popularité (infos supplémentaires):** Points to the number of fans '40 Fans' for the user 'XenoPhil' in the third post.
- Relation structurelle (réponds à):** Points to the 'Répondre' button of the second post, indicating it is a reply to the first post.

The forum posts include:

- Post 1:** User 'nicolas22' (0 Fans) posted 'il y a 54 minutes (11h47)'. Content: 'pour avoir des réponses concernant l'espionnage de l'Europe il suffit de demander a nos amis anglais, ils sont copain comme cochon .Il faut que les anglais sortent de l'Europe'.
- Post 2:** User 'Isabelle Forger' (27 Fans) posted 'il y a 35 minutes (12h06)'. Content: 'ou qu'ils choisissent leur camp...'.
- Post 3:** User 'SUPER UTILISATEUR DU HUFFPOST XenoPhil' (40 Fans) posted 'il y a 1 heure (11h23)'. Content: 'Video Not Available This video has not been made available in your country by the owner' Ils devraient prendre exemple sur les "joumaux", à la NSA ??? (au moins eux ils savent garder les infos secrètes)'.
- Post 4:** User 'pablico' (168 Fans) posted 'il y a 2 heures (11h01)'. Content: 'Écouter aux portes, et regarder par les trous de serrures est ce du terrorisme?'.
- Post 5:** User 'SUPER UTILISATEUR DU HUFFPOST XenoPhil' (40 Fans) posted 'il y a 1 heure (11h36)'. Content: 'Lorsque c'est les "ncains", juste derrière il y a les drones qui assassinent !!!! Alors OUI, dans ce cas ce sont bien des terroristes ... D'ailleurs, rien d'étonnant à cela, les USA sont ou ont été derrière quasiment tous les terroristes comme BenLaden, les rebelles syriens, le Mossad, Jundollah en Iran etc etc'.

Objectif général :

deux types d'utilisateurs

l'analyste des forums : comprendre les discussions entre les utilisateurs et leurs thématiques

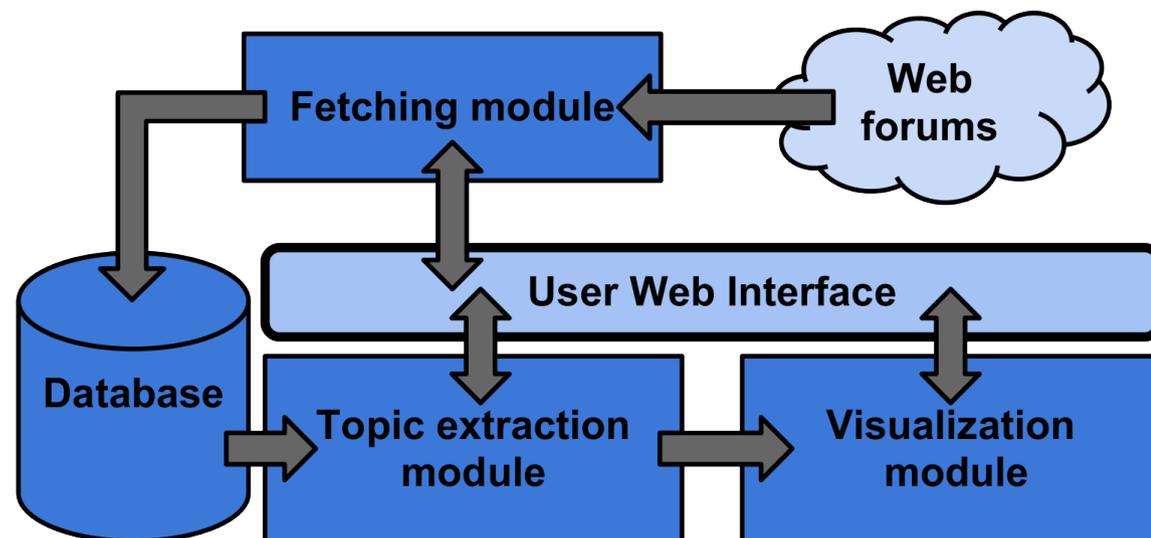
le chercheur : construire des jeux de données forums, analyser les évolutions des thématiques de discussion

Notre proposition : CommentWatcher

Plateforme Web opensource (GPLv3)

4 tâches :

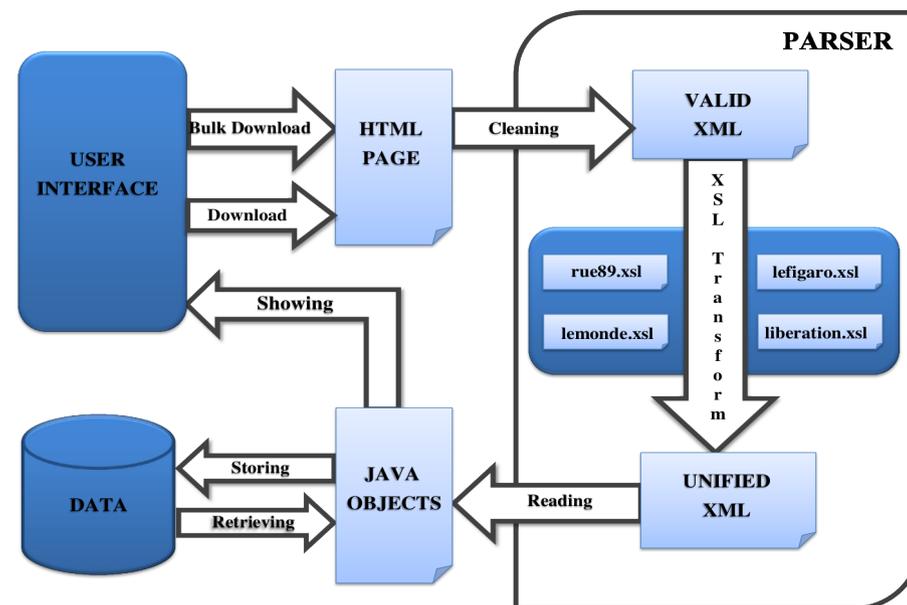
- Récupération des données à partir d'Internet
- Extraction de thématiques
- Visualisation de thématiques comme un nuage d'expressions et l'évolution temporelle
- Visualisation du réseau social sous-jacent



Module I. Récupération des données

Récupérer le texte des discussion sur des forum en ligne, ainsi que des méta-données sur les utilisateurs et leur relations (e.g., pseudo, nom, date, qui répond à qui *etc.*)

- Méta-parseur, indépendant de la structure des pages web
- Support pour de nouveaux sites via des fichiers de définition
- Recherche des forums supportés via une requête, en utilisant l'API Bing
- Téléchargement « en masse » des forums



Module II. Extraction des thématiques

Extraire les thématiques de discussions, à partir d'un sous-ensemble de documents, en utilisant des algorithmes issues de data mining.

3 algorithmes supportés :

- Topical Ngrams (suite Mallet [\[MCC02\]](#))
- CKP [\[RIZ10\]](#)
- Dynamic Topic Models [\[BLE06\]](#)
(en développement)

```

---> Partitions:
---> Cluster 0 [101]: .....
---> Cluster 1 [90]: .....
---> Cluster 2 [128]: ..... texte_81 .....
---> Cluster 3 [192]: ..... texte_81 ....

```

Result - Cluster description:

```

-> Centroid[0]: "jours fériés"
-> Centroid[1]: "travailler plus pour gagner plus"
-> Centroid[2]: "commémoration"
-> Centroid[3]: "histoire de france"

```

Exemple de sortie du logiciel d'extraction de thématiques, inclus dans **CommentWatcher**

Démonstration Vidéo

Video demonstration

CommentWatcher: An open source web-based platform for analyzing discussions on web forums

Marian-Andrei Rizoïu, Adrien Guille, Julien Velcin
ERIC Lab - Université Lumière Lyon 2
Université de Lyon, France



Site de Présentation : <http://mediamining.univ-lyon2.fr/commentwatcher>

Conclusion

- plateforme Web opensource
- parseur facilement adaptable aux changements de structure des sites
- visualiseurs qui permettent la compréhension rapide des thématiques de discussion et la structure du réseau social sous-jacent

Développements futurs

- extraction de thématiques temporelles et visualisation adaptée
- intégration du calcul des mesures pour les réseaux sociaux
- évoluer la visualisation côté client (applet) vers une visualisation côté serveur
- intégrer l'évaluation de la cohésion sémantique des thématiques [\[MUS11\]](#)

Équipe de développement

Développeur principal

Marian-Andrei RizoIU

Développeurs

Mouhamadou Bamba Kane (Master 2)

Brian Ampwera (Master 1 DMKM)

Cyril Briquet (Master 1 Informatique)

Nikolay Anokhin (Master 1 DMKM)

Supervision

Marian-Andrei RizoIU, Julien Velcin, Adrien Guille

Projets de recherche

ImagiWeb

CRTT-ERIC

ERIC-ELICO

Conversession

Je vous remercie pour votre attention !

Site de présentation : <http://mediamining.univ-lyon2.fr/commentwatcher>

Installation publique : <http://mediamining.univ-lyon2.fr:8080/CommentWatcher>

Bibliographie

- [**AME12**] S. Amer-Yahia, S. Anjum, A. Ghenai, A. Siddique, S. Abbar, S. Madden, A. Marcus, and M. El-Haddad. Maqsa: a system for social analytics on news. In SIGMOD '12, pages 653–656, 2012.
- [**GUI13**] A. Guille, C. Favre, H. Hacid, and D. Zighed. Sondy: An open source platform for social dynamics mining and analysis. In SIGMOD '13, 2013.
- [**MCC02**] A. K. McCallum. Mallet: A machine learning for language toolkit.
<http://mallet.cs.umass.edu>, 2002.
- [**RIZ10**] M.-A. Rizoiu, J. Velcin, and J.-H. Chauchat. Regrouper les données textuelles et nommer les groupes à l'aide des classes recouvrantes. In EGC '10, page 561, 2010.
- [**BLE06**] David M Blei and John D Lafferty. Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning, pages 113–120. ACM, 2006.
- [**MUS11**] Claudiu Musat, Julien Velcin, Stefan Trausan-Matu and Marian-Andrei Rizoiu. Improving topic evaluation using conceptual knowledge. In International Joint Conference on Artificial Intelligence, Proceedings of the Twenty-Second, volume 3 of IJCAI 2011, pages 1866–1871. AAAI Press, 2011.