

## *The Observatorium*

### **Observation et analyse de réseaux de communication a grande échelle**

Jorge Louçã et David Rodrigues

Lisbon University Institute et LabMAg - Laboratory of Agent Modelling

#### **1. Introduction**

La notion d'échelle, essentielle dans l'étude des systèmes sociaux humains, est le point de départ de cette communication, qui prétend discuter et démontrer quels sont les résultats que nous pouvons espérer de l'analyse de grandes quantités de données de communication, à l'échelle globale, ceci en ayant recours à des propositions méthodologiques et à des outils spécifiques,

L'Internet est devenu le media le plus important pour les systèmes sociaux qui interactive-ment s'échangent des idées, et génèrent de la connaissance dans le monde entier. Des réseaux sociaux et de communication sont utilisés pour informer, échanger des arguments, former des opinions concernant la politique, l'économie ou la culture. La génération de connaissances est supportée par des blogs, des nouvelles, des articles d'opinion écrits par des journalistes, des politiciens, des scientifiques et autres. Ces systèmes sont très dynamiques et complexes, étant donné leur interconnexion et interdépendance, évoluant de façon distribuée.

Le projet *Observatorium*<sup>1</sup> (Rodrigues et Louçã, 2010) cherche à comprendre les dynamiques dans les systèmes de génération de connaissances et d'opinions. Ces systèmes sont supportés par des grands réseaux de communication. Pour comprendre ces dynamiques, l'*Observatorium* propose la surveillance en temps réel de la structure et des relations entre des topiques de discussion qui fluent sur Internet.

---

<sup>1</sup> <http://theobservatorium.eu>

Suite à une brève référence sur la récolte et la collection de données, le texte présente une approche pour le traitement de données issus de réseaux de communication, à travers un ensemble de propositions pour la détection de topiques dans les journaux on-line. La caractéristique principale de ces propositions est son indépendance sémantique par rapport au langage utilisé dans la communication, permettant la détection de topiques et la caractérisation de la dynamique des réseaux de communication sans avoir recours à des outils d'analyse sémantique normalement utilisés. En effet, l'utilisation de dictionnaires et d'ontologies prétend palier la difficulté de comprendre les spécificités du langage utilisé. La catégorisation automatique de texte proposée par la linguistique, par le traitement de langage naturel et par la statistique ont conduit au développement de différentes approches, tels que des modèles de régression, des approches Bayésiennes, la classification du voisin le plus proche, les réseaux neuronaux ou le clustering automatique (Cachopo and Oliveira, 2003; Miao and Qiu, 2010; Sole et al., 2010). Cependant, l'utilisation systématique de ces outils est très couteuse en termes de ressources nécessaires pour la computation de grandes quantités de données. La grande majorité des ces méthodes est supervisée. Celles-ci nécessitent d'un ensemble de données pour entraîner le système, où des documents classifiés par des humains sont utilisés tant qu'input pour l'apprentissage initial de la méthode utilisée. D'un côté l'utilisation de méthodes d'analyse dépendant du langage humain, et d'un autre côté le besoin de classification par des spécialistes, font que l'application de méthodes traditionnelles ne soit pas recommandée pour l'étude de grandes quantités de données, dynamiques au cours du temps. Ceci justifie nos propositions, basées sur une analyse non-sémantique de la communication.

## **2. Récolte et collection de données**

L'*Observatorium* récolte et collectionne actuellement les journaux en ligne "Público" et "Jornal de Negócios" (Portugal), "The Times" et "The Guardian" (UK), "The Australian" (Australia), "Le Monde" (France), "El País" (Espagne), "nol.hu" (Hongrie), "Corriere della sera" (Italia), "Irish Times" (Irlande) et "Politika" (Bulgarie).



Figure 1 : Exemples de journaux en ligne récoltés et collectionnés par l’*Observatorium*

La récolte et la collection d’autres ressources en ligne débiteront probablement au cours de l’année de 2011, en incluant une diversification des ressources au delà des articles de journaux.

### 3. Surveillance de la dynamique de sujets dans les media en ligne

La première expérience réalisée dans le cadre de l’*Observatorium* a eu comme objectif la surveillance de la dynamique de sujets issus de journaux en ligne. Une méthodologie générale pour l’acquisition et le traitement de textes issus de l’Internet à été initialement définie. Cette méthodologie s’initie par l’obtention de documents web (par exemple des pages HTML), ensuite elle traite ces documents web pour en extraire la partie du texte la plus importante, puis finalement elle classe les éléments textuels obtenus de façon à représenter graphiquement ces éléments et leurs relations, sous la forme d’un réseau (voir Figure 2).

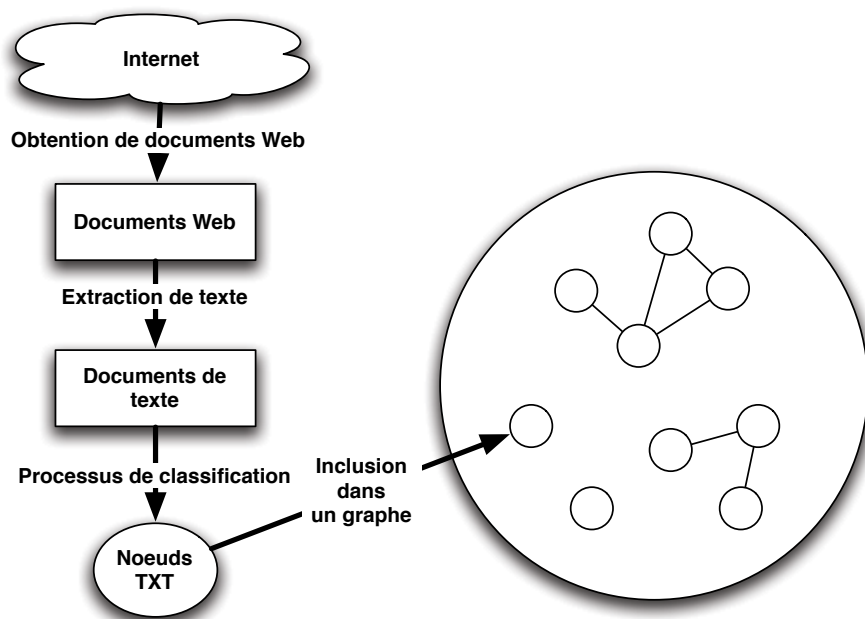


Figure 2 : Méthodologie générale pour l’acquisition et le traitement de textes issus d’Internet

Cette méthodologie générale permet la classification d’articles, la représentation de réseaux de sujets, et enfin la surveillance de ces réseaux en temps réel. Ce processus est par la suite détaillé.

### 3.1. Extraction de texte à partir de documents web

L’extraction de texte a été réalisée à partir de fichiers HTML, avec l’utilisation de l’algorithme text-to-tag ratio (TTR) proposé par Weninger (2008). TTR est une heuristique pour extraire le contenu de pages HTML. Il s’agit du calcul, pour chaque ligne de code HTML, du ratio de la somme des caractères non-HTML par la somme des caractères de tags HTML. Initialement les scripts et les annotations sont éliminés, aussi bien que les lignes vides. Un fort taux de TTR indique la section du fichier contenant le texte important, c’est-à-dire le texte d’un article expurgé de toutes tags HTML.

Exemple:

*Considérons l’extraction du contenu d’une page du journal bulgare “Politika”.*





Figure 5 : Résultat de l'extraction du texte à travers l'application du TTR

L'application de l'algorithme TTR s'est montrée particulièrement efficace pour identifier les articles de journaux en ligne. La phase suivante a consisté à analyser les ensembles d'articles permettant d'identifier les sujets traités par les journaux.

### 3.2. Composition du réseau d'articles

La méthodologie d'identification de sujets, sans avoir recours à aucune analyse sémantique des textes, a considéré initialement que chaque article est constitué par un ensemble de mots auquel a été retiré toute ponctuation et les mots jusqu'à trois lettres. Chaque ensemble de mots est représenté par un nœud et additionné à un réseau, où le nœud est lié à d'autres nœuds en fonction de sa distance envers les nœuds du réseau. Un réseau de nœuds est donc obtenu, chacun représentant un article d'un journal. Des clusters dans le réseau peuvent alors être observés, qui représentent un sujet traité par plusieurs articles.

La distance entre les nœuds est calculée en ayant recours à l'*indice de Jacard*. Ce coefficient de similarité est une mesure statistique fréquemment utilisée pour comparer la similitude et la diversité d'ensembles de données. L'*indice de Jacard* est défini comme la dimension de l'intersection divisée par la dimension de l'union de deux ensembles de données.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Expression 1: L'indice de Jacard

Considérons la mesure de la distance des articles. Si la distance entre deux nœuds est inférieure à une limite donnée, alors les nœuds sont liés. Cela veut dire que si la mesure entre deux ensembles de mots est inférieure à une limite, alors ces deux ensembles de mots sont liés dans le réseau et donc ils appartiennent au même sujet.

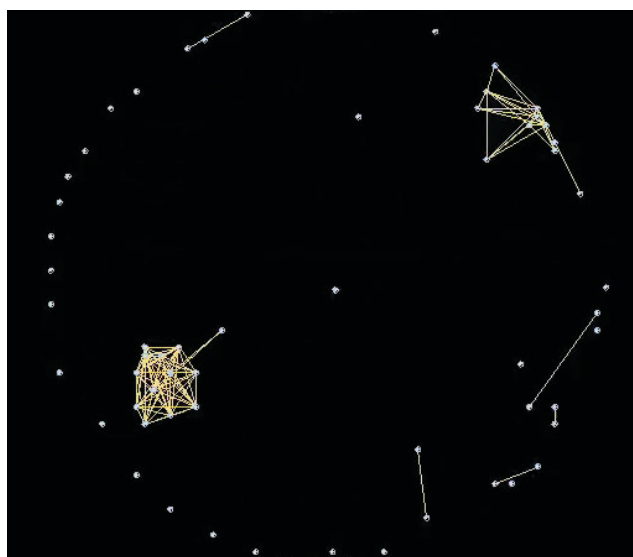


Figure 6 : Exemple de réseau d'articles avec deux sujets remarquables

### 3.3. Surveillance de la dynamique d'un réseau d'articles

La surveillance de la dynamique d'un réseau d'articles a été réalisée à travers l'application du concept de time to live (TTL). Le temps de vie d'un nœud du réseau est donné par son TTL. Chaque fois qu'un nœud est additionné au réseau, le TTL des nœuds qui établissent une nouvelle connexion avec lui est augmenté, et le TTL des autres qui ne se connectent pas est diminué. Ce mécanisme permet d'éliminer régulièrement les nœuds anciens, qui ne reçoivent plus de nouvelles connexions pendant une période de temps déterminée, comme cela arrive, par exemple, aux nœuds isolés dans le réseau.

### 3.4. Détection de sujets

Enfin, la détection de sujets dans le réseau dynamique d'articles est réalisé en ayant recours au concept de *variation d'information* (VI), proposé par Meilã (2007). VI est une métrique qui permet d'identifier des ensembles de documents similaires, à travers la mesure des distances entre clusters. VI utilise le concept d'entropie pour mesurer la quantité d'information perdue ou gagnée au moment du passage d'une version du réseau pour une autre actualisée<sup>2</sup>.

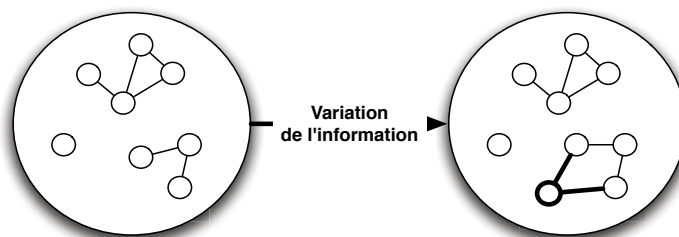


Figure 7 : Concept de variation d'information de Meilã (2007) appliqué à un graphe générique

Nos expériences ont utilisé les paramètres  $TTL = 100$ ,  $j_{min} = 0.5$  et  $VI_{min} = 0.5$ . Chaque interaction a calculé la valeur de VI d'une version du réseau pour la version suivante, ce qui a permis d'identifier les moments précis où VI a dépassé  $VI_{min}$ . Ces moments correspondent à l'élimination de parties significatives du réseau, suite à l'application du TTL aux nœuds anciens. Cela a rendu possible la détection du moment d'élimination de clusters de dimension significative, indiquant l'existence de sujets dans des ensembles d'articles.

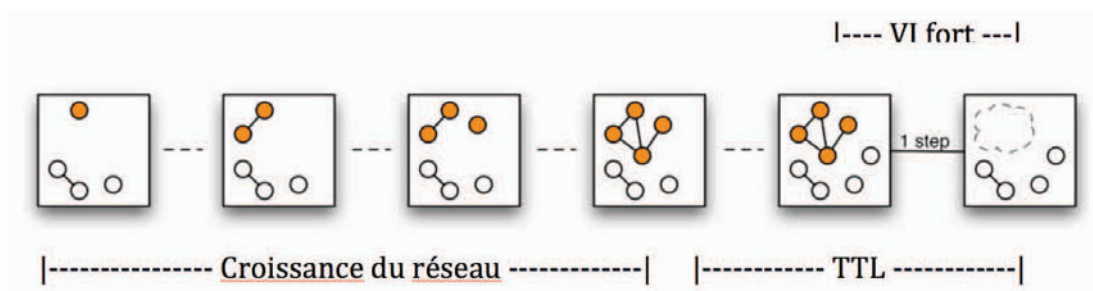


Figure 8 : Représentation schématique de la croissance et détection d'un sujet à travers VI

<sup>2</sup> Pour une discussion de l'application de la *variation de l'information* de Meilã aux medias voir aussi (Rodrigues et Louçã, 2010 et 2009)

L'expérimentation des mécanismes décrits pour la récolte de documents web, le traitement de texte significatif et l'identification de la dynamique des sujets en journaux en ligne a été réalisé sur des données récoltées par l'*Observatorium*. Cette expérimentation est synthétiquement décrite dans la section suivante.

## 4. Exemple

7928 articles ont été obtenus à partir du journal "Público" en ligne, du 11 Novembre 2009 au 25 Janvier 2010. Les textes des articles ont été extraits des pages HTML de ces articles.

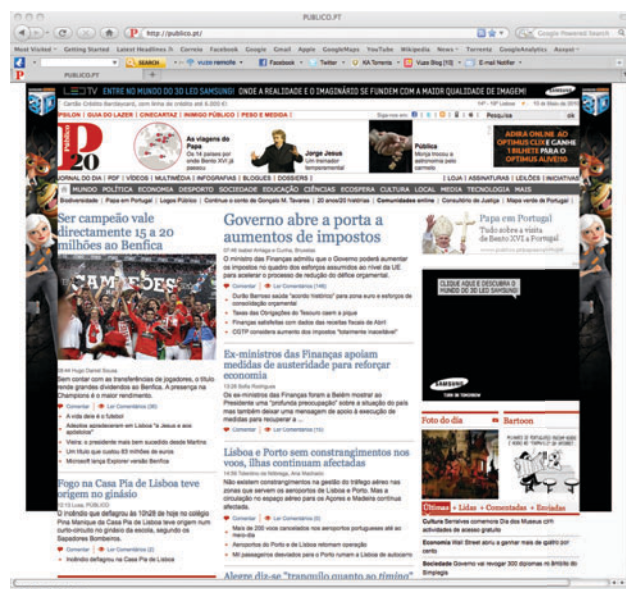


Figure 9 : Le journal "Público" en ligne

La variation d'information a été calculée par la suite pour la période en référence. Les valeurs de VI obtenues sont représentées dans la Figure 10.

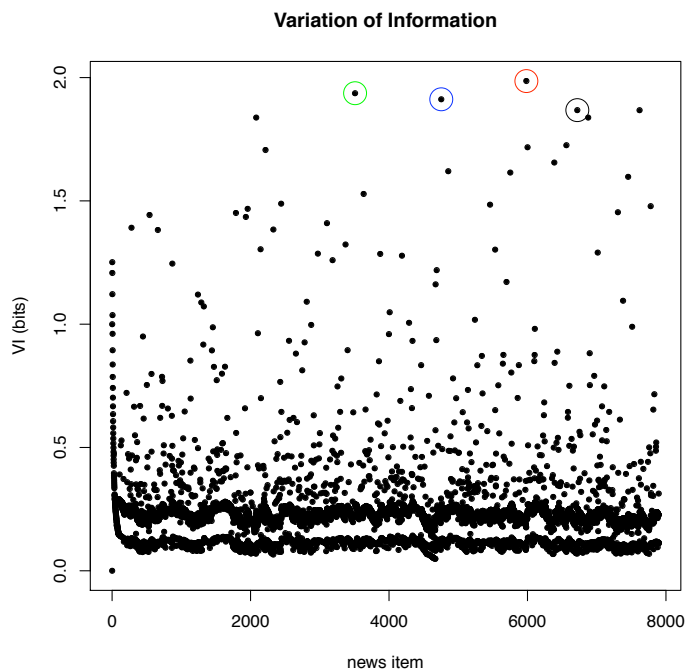


Figure 10 : VI correspondant à la période du 11 Novembre 2009 au 25 Janvier 2010

Les valeurs de VI les plus hautes sont signalées en couleur. Ces valeurs représentent les changements les plus importants observés dans la structure du réseau d'articles. Ces changements sont le résultat de l'élimination de clusters entiers d'articles, ce qui permet la détection de sujets dans l'ensemble de données. Les sujets peuvent, par exemple, être visualisés dans un état du réseau:

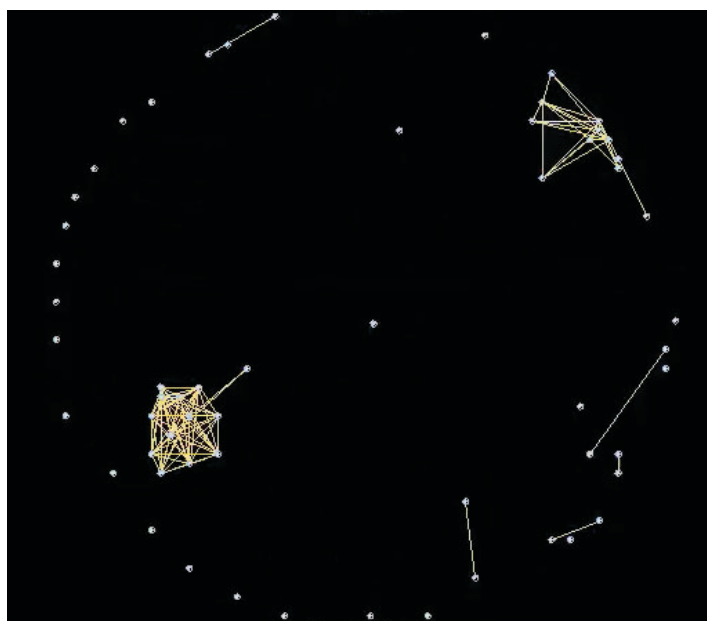


Figure 11: Exemple d'un état du réseau d'articles avec deux sujets remarquables

Suite à sa détection, les sujets sont caractérisés de la façon suivante: chaque cluster, composé par plusieurs articles, est traité en tant que texte unique et continu. Un histogramme compare la fréquence de ses mots. Les mots les plus fréquents caractérisent chaque sujet. Par exemple, trois sujets ont été caractérisés par:

“paiement” “international” “bourse” “étude” “négociation” “commentaires”  
“suivant” “commentaires” “industrielle” “cent” “confiance” “exportations”  
“Madrid” “tournoi” “Sporting” “joueur de football” “joueurs” “jeux” “commentaires”

La représentation synthétique des sujets et de leur dynamique, permettront éventuellement d’autres recherches par des équipes multidisciplinaires.

## 5. Perspectives

Les lignes de recherche actuellement en développement dans le cadre de l’*Observatorium* sont caractérisées par les objectifs suivants:

Développement d’une bibliothèque informatique d’outils crawler, non uniquement pour les appliquer a des journaux en ligne, mais aussi pour d’autres types de données représentatifs de la communication sur des grandes réseaux de communication, tels que les blogs, twitter, facebook, etc;

Développement d’une bibliothèque d’outils algorithmiques pour la découverte de corrélations entre des données issues de différentes origines dans “l’espace web”;

Mise à la disposition de données pour la communauté scientifique;

Mise à la disposition d’outils en ligne permettant la visualisation de la dynamique des réseaux de communication en temps réel;

Mise à la disposition de notre bibliothèque de software pour la recherche, sur une licence open source, accessible à partir de l’adresse <http://theobservatorium.eu/>

## 6. Bibliographie

- Cachopo, A.M.D.J.C., Oliveira, A.L.: An Empirical Comparison of Text Categorization Methods. *String Processing and Information Retrieval* (2003) 183–196
- Meilă, M.: Comparing clusterings- an information-based distance. *Journal of Multivariate Analysis* 98 (2007) 895
- Miao, Y., Qiu, X.: Hierarchical Centroid-based Classifier for Large Scale Text Classification. (2010) 3–6
- Solé, R.V., Corominas-murtra, B., Valverde, S., Steels, Luc.: Language Networks: Their Structure, Function, and Evolution. *Complexity* 00 (2010) 1–7
- Weninger, T., Hsu, W.H.: Text Extraction from the Web via Text-to-Tag Ratio. 2008 19th International Conference on Database and Expert Systems Applications (2008) 23–28
- Rodrigues, D.: The structure of news: monitoring online media topics with mutual information. *European Conference on Complex Systems, Portugal* (2010)
- Rodrigues, D., Louçã, J.: Mutual information to assess structural properties in dynamic networks. *European Conference on Complex Systems, UK* (2009)
- Rodrigues, D., Louçã, J.: The Observatorium: monitoring topic trends in online media. *Second Annual Meeting of the COST Action MP0801 "Physics of Competition and Conflicts"*, Bulgaria (2010)