Diversité de l'Information dans les Sites de Presse

Cyril Laitang, Elöd Egyed-Zsigmond, Sylvie Calabretto

Université de Lyon
LIRIS, INSA de Lyon

Mots-clés : catégorisation, web, sémantique, presse, flux RSS, thématique, ontologie

Keywords: RSS feeds, web, semantic, categorization, press, ontology

Résumé : La multiplication des acteurs de communication sur internet devrait logiquement entrainer une plus grande diversité dans les sources d'information disponibles. Toutefois de plus en plus d'études récentes mettent en doute cette hypothèse. Le projet IPRI a été créé afin de faire le point sur ces croyances et déterminer les schémas de propagation des actualités sur ce media. Notre principal objectif est de fournir une aide à l'analyse des informations de presse sur des évènements d'actualités. Nous assistons la catégorisation en fournissant un outil de classification dynamique des thématiques et des sujets extraits de flux RSS au moyen de ressources du Web sémantique. Notre solution lie la catégorisation sur des domaines non spécialisés à des sources et concepts sémantiques tel que des ontologies et des thésaurus.

Abstract: As internet communication actors grow in number it is widely admitted that informational offer diversity over this media should increase as well. However we recently came over more and more studies that seem to go against that assumption. The IPRI project was created in order to asset these believes and to figure out news events spreading schemes over the web. Our main objective is to help press information analysis over current events. We assist categorization by providing dynamic thematic and subject sorting tool over RSS feeds by the mean of semantic web resources. Our solution links information categorization over unspecialized data sources and various fields with semantic concepts like thesaurus and ontologies.

1 Introduction

1.1 Problématique

Considéré comme le socle de la démocratie, le pluralisme de l'information est l'objet d'une régulation dans les médias écrits et audiovisuels. Sur Internet en revanche il est permis de penser que la multiplicité des sources accessibles garantit naturellement cette diversité de l'information.
Qu'en est-il réellement? C'est pour répondre à cette problématique qu'a été créé le projet IPRI [1]. En partenariat avec des sociologues et des journalistes nous cherchons à aider la catégorisation des articles de presse depuis un échantillon représentatif des grandes familles de type de publication. Ceci afin de déterminer le niveau de répétitivité et le schéma de propagation de l'information.
Dans ce papier nous décrirons comment l'ajout et l'utilisation d'informations sémantiques peut fournir de l'aide à la fois à la catégorisation thématique et au rapprochement des sujets traités sur un intervalle temporel défini.

1.2 Objectifs

Notre problématique tente de répondre à deux besoins de catégorisation dont le sens propre au journalisme mérite une redéfinition :
La thématique, statique elle concerne le domaine dans lequel se classe l'article. Sport, politique, etc. Elle est nécessaire à toute catégorisation ultérieure dans la mesure où elle donne des informations utiles au regroupement par sujet. Il est à noter également que malgré sa normalisation par de grands organes de presse elle n'est peu ou pas appliquée à la publication de la grande majorité de nos sources (blogs, agrégateurs). Exemple: thématique sport, politique, etc.
Le sujet, ce dernier apparaît, évolue et disparaît au cours du temps. Cela implique deux problématiques : le choix de la distance temporelle entre deux analyses et son regroupement. Exemples : élection américaine, tremblement de terre, nouvelles lois.
Les deux types de catégorisations mentionnées ci-dessus sont, à l'heure actuelle, effectuées manuellement : par une annotation de l'article à sa publication pour la thématique; et par une analyse subjective de journalistes, sociologues, ou analystes de presse pour le regroupement par sujet. L'exemple de l'étude sur les sujets de presse [7] illustre les besoins existants dans ces domaines. En effet, cette analyse a demandé trois jours

[1] Internet, Pluralisme et Redondance de l'Information. Projet soutenu par la Maison des Sciences de l'Homme-Paris Nord

d'annotation manuelle, durée que notre solution permettrait de réduire considérablement.
Notre solution se caractérise par une série d'apports et la combinaison d'approches ayant déjà fait leurs preuves dans les domaines concernant à la fois la catégorisation et l'enrichissement sémantique. C'est pourquoi, dans la suite de ce papier, nous commencerons par un bref état de l'art des connaissances actuelles en RI[2] orienté catégorisation, avant de poursuivre par une série de définitions accompagnées d'exemples sur les sources sémantiques utilisées par notre solution. Par la suite nous décrirons les apports de notre solution pour l'aide à la catégorisation et à l'analyse des dépêches de presse. Enfin nous présenterons brièvement l'état actuel de notre prototype et des résultats obtenus avant de conclure sur les perspectives ouvertes à notre solution et les implémentations à venir.

2 État de l'art

Du fait du sujet même de notre problématique de recherche (la catégorisation assistée des dépêches de presse) il nous semble important dans un premier temps de rappeler et d'expliciter les concepts fondamentaux en RI orienté sur ce domaine. Nous définirons donc dans un premier temps les concepts de RI associés à notre projet avant d'approfondir la catégorisation à proprement parler. Par la suite, du fait de l'intégration de multiples sources sémantiques, il nous semble nécessaire de fournir un descriptif succinct ainsi qu'une présentation des sources utilisées.

2.1 Recherche d'information et catégorisation

On définit la Recherche d'Information par « l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents ceux qui sont susceptibles de répondre aux besoins de l'utilisateur » [1] [2]
La catégorisation consistant en un rapprochement d'éléments similaires, il existe une forte corrélation entre les techniques dites de RI et celles associées à notre domaine d'étude. En effet, dans les deux cas il s'agira de déterminer la pertinence d'un ensemble, soit par rapport à une requête, soit par rapport à des documents voisins. C'est pourquoi le premier des deux processus clef de RI, à savoir le processus de « représentation », également appelé processus d'indexation. [1] [2] présente un fort intérêt pour notre système de catégorisation.

[2] Recherche d'Information

2.1.1 Indexation

Les processus d'indexation consistent en la description d'un document à l'aide de la représentation du contenu de celui ci. [1] [2] Autrement dit ils se caractérisent par la sélection puis par la pondération des termes pertinents.
La sélection des descripteurs se fait au moyen d'un équilibrage entre deux méthodes que sont la discrimination (distinction avec le reste du corpus de document) et la représentation (caractérisée par le contenu, modélisation du sujet dont traite le document.)

2.1.2 Extraction

Les processus d'extraction comportent deux principales familles d'approche :

- *L'extraction par cooccurrence*: est l'utilisation de groupes de mots pour définir un concept, elle mesure le nombre d'apparitions d'un mot sémantiquement non vide et son positionnement dans le document.
- *L'extraction par analyse linguistique*: (détection de patrons de mots NOM PREP. NOM) qui donne des informations sur la syntaxe et l'importance du terme. (un nom sera plus important qu'un adjectif par exemple).

Au regard des études sur le sujet et des solutions développées jusqu'ici l'approche dite mixte, c'est à dire prenant en compte à la fois le comptage des occurrences du terme dans le document et sa forme lexicale semble la plus appropriée à notre projet. Nous utilisons *TreeTagger*[3] [12] qui permet à la fois l'extraction, le traitement (transformation des verbes à l'infinitif, élimination des formes plurielles et autres problèmes pouvant nuire à la performance des comparaisons de termes) et l'identification de la famille grammaticale du terme.

2.1.3 Distance sémantique

Afin de répondre à notre problématique nous cherchons à effectuer deux types de rapprochements : un rapprochement entre les flux et les thématiques statiques (explicitées par notre thésaurus extrait des catégories normalisées de l'*IPTC*[4] , mais nous y reviendrons dans la section contribution) et un rapprochement sémantique entre les articles de presse eux même. Il nous semble donc que les mesures dites de « *similarité sémantique* » sont les plus adaptées à notre problématique.

[3] "TreeTagger" http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[4] Information Press Telecomunication Council, (voir état de l'art thésaurus).

Il convient donc de rappeler brièvement ce que sont ces mesures. En premier lieu il faut faire les distinctions entre la similarité sémantique et la proximité sémantique [1]. La similarité peut se calculer soit par la distance entre les arcs d'une représentation arborescente des termes, soit par un comparatif entre l'occurrence des termes. La proximité quant à elle prend en compte les relations entre les éléments que nous représentons dans notre solution sous forme de graphes, les relations étant les arcs et les éléments les nœuds.
Du fait de la structure même de nos sources de données au sein de notre prototype (se reporter à la section prototype de ce papier) nous utiliserons à la fois un système de calcul basé sur l'occurrence des termes et la distance entre les concepts en suivant le plus court chemin qui les sépare.

2.2 Les ressources sémantiques

Les travaux récents en recherche d'information et en catégorisation tendent vers la prise en compte de la sémantique. Pour rappel, la sémantique est l'étude du sens des mots et des relations qui les lient. L'analyse de corpus peut présenter de nombreux problèmes auxquels l'ajout de sources sémantiques peut pallier efficacement par des procédés tels que :

- *La désambiguisation* : des mots structurellement proches peuvent voir le sens qui leur est attribué changer du tout au tout (on parle alors de mots appartenant à plusieurs catégories syntaxiques). Ainsi par l'utilisation d'ontologies et des procédés de calcul de distance sémantique on peut ré-identifier le sens de ce mot.

- *L'enrichissement* : en associant les synonymes et les termes sémantiquement proches on précise le domaine de la recherche de correspondance entre nos termes et leurs catégories.

La distinction entre *Ontologies*, *Thésaurus* et *Bases de Données lexicales* est ténue dans la littérature [4]et parfois même dans la définition qu'en donnent leurs créateurs. Nous avons donc tenté d'en redonner une définition simplifiée dans la suite de ce papier tout en les illustrant par certains de leurs représentants utilisés dans notre prototype.

2.2.1 Thésaurus

Les thésaurus peuvent être définis comme des « *Ensembles hiérarchiques de termes clés représentant des concepts d'un domaine particulier.* »[5] Généralement organisés de façon thématique, les éléments de leur vocabulaire sont liés entre eux par des liens sémantiques qui peuvent être

[5] "Thesaurus." Dicodunet http://www.dicodunet.com/definitions/moteurs-de-recherche/thesaurus.htm

: la synonymie, l'équivalence, la spécificité (lien vers un concept de sens plus précis), ou la généralisation (lien vers un concept de sens plus large).
Dans le cadre de notre projet nous avons effectué l'étude d'un large échantillon de thesauri francophones librement exploitables tels qu' *Agrovoc*, *Jurivoc, Hurivoc, Dadi, Delphe, GeoEthno* ou encore *Eurovoc*.
Après l'analyse de soixante-quinze d'entre eux nous avons découvert qu'environ la moitié mettent à disposition leur ontologie en téléchargement et que sur cette moitié restante seulement cinq l'étaient sous la forme d'un ensemble de documents normalisés et réellement exploitable. La majorité se limitant à une série de *pdf* commentés. Concernant le volume lui même oscille entre quelques dizaines de milliers et plusieurs centaines de milliers.
La spécificité même des thésaurus associés aux particularités de notre problématique (documents francophones, domaines variés) font que l'intégration à notre base des seuls thésaurus disponibles risquait de parasiter le processus de calcul de la thématique. Nous avons donc fait le choix de n'intégrer que le thésaurus traduit par *Raphael Troncy* [4] des catégorisations de dépêches de presse de l' *IPTC*[6] et de l'utiliser comme base de référence thématique statique.
Pour information, le thésaurus de l'*IPTC* contient 1400 catégories, réparties sur trois niveaux d'abstraction et décrites succinctement.

2.2.2 Ontologies

On définit les ontologies comme des « Ensembles structurés des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances». La différence la plus intéressante par rapport à un thésaurus est qu'une ontologie offre un plus large champ de relations entre ses concepts que la simple hiérarchie comme par exemple la possibilité de retournement soit la navigabilité bidirectionnelle des arcs.
Parmi les ontologies disponibles nous avons porté notre choix et nos analyses sur deux d'entre elles:

- *DBPedia* : ontologie multi-domaine et multilingue créée entre autre à partir des « *infobox* » de *Wikipedia*. Son utilisation à l'heure actuelle reste peut répandue, même si elle a déjà été adoptée comme source de données sémantiques dans plusieurs projets de recherche et d'ingénierie. [4] [6]. Pour information l'ontologie *DBPedia* regroupe 2,6 millions de concepts dans 36 langues.

- *Yago* [10][11] : une ontologie généraliste anglophone basée sur *Wordnet* et *Wikipedia*. Son niveau de généralisation le plus élevé est établi d'après le système de classification de *Wikipedia*. Son intérêt

6 The IPTC-NAA standards."
http://www.iptc.org/cms/site/index.html?channel=CH0086

principal résidant en son système de représentation des liens entre concepts sous la forme {Sujet}-{Relation}-{Sujet}. *Yago* regroupe 95% des concepts *Wikipedia*.

Ces ontologies sont librement exploitables au format RDF, par « *endpoint* » (requête sur service web) et interrogeable via *SPARQL*. Il est à noter que *Yago* fournit un outil java de traitement et de conversion que nous avons utilisé pour franciser ses concepts par requête *SPARQL*.

2.2.3 Bases de données lexicales

Nous définissons les bases de données lexicales comme des ensembles de termes liés par synonymie ou proximité sémantique et organisés hiérarchiquement.
Au contraire des thésaurus, les bases de données lexicales sont généralistes. Et à la différence des ontologies les concepts exprimés sont simplifiés. Bien que considéré par beaucoup comme un thésaurus, *WordNet*[7] est l'illustration parfaite de ce que nous qualifierions de base de données lexicale.
Le problème de ces sources de données sémantiques est que leur représentant le plus significatif est en anglais, et du coup non exploitable par notre solution. Nous sommes donc en phase de recherche d'une base de données lexicale francophone récente et complète.

3 Contribution

Afin d'étudier le pluralisme des informations publiées sur le web, nous voulions analyser les informations publiées par les différents sites d'information, qu'ils soient des sites web de journaux classiques, des webzines (journaux exclusivement numériques), des blogs, des sites participatifs, des portails, des agrégateurs d'information ou des agences de presse. Une première idée était de prendre en compte le contenu complet de ces pages web, mais devant l'ampleur de la tâche face aux ressources disponibles, nous avons limité notre champ aux sites dotés d'un flux RSS. Un tel choix présente l'avantage de traiter des données homogènes issues de sources différentes.
Nous avons recueilli 89 flux RSS répartis dans les catégories ci-dessus. Nous les avons enregistré en continu pendant plusieurs semaines. Pour chaque item émis par un flux RSS, nous avons recueilli son titre, la date et l'heure d'émission, …
Nous proposons de raffiner le modèle statistique de calcul de la proximité des catégories de sujet au moyen de thésaurus, d'ontologies et de

[7] WordNet - Princeton University Cognitive Science Laboratory." http://wordnet.princeton.edu/

dictionnaires lexicaux. Parallèlement nous proposons des méthodes d'identification, d'extraction et de pondération des termes significatifs. Enfin nous ouvrons des perspectives de développement de notre projet quant à la catégorisation temporelle des dépêches de presse alternativement nommée « *Catégorisation de sujets* ».
Le schéma ci-dessous représente la structure de la base de données dont il sera question dans les sections suivantes. A titre d'information (même si il en sera question plus longuement dans les sections consacrées à l'intégration des sources sémantiques et à l'agrégation) : les flux RSS agrégés sont répartis sur les tables *lemmes*, *rss_item* ; les thématiques et leur relation hiérarchique sont contenues dans les tables *thema_lemmes*, *thematique*, *relation*, et *thesaurus*. Les ressources *DBPedia* dans *dbpedia_ressources* ; et les informations sur les sources de nos flux dans *tags*, *fluxrss_tags*, *fluxrss*.

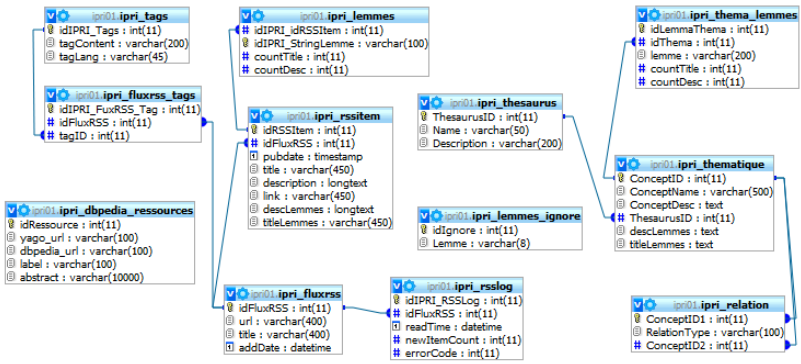

Figure 1 Structure de la base de données du projet

3.1 Intégration des sources sémantiques

De notre section état de l'art nous avons choisi d'intégrer à l'heure actuelle trois sources sémantiques, à savoir : le thésaurus de l'*ITPC*, l'ontologie *DBPedia* et l'ontologie *Yago*. Il est à noter également que nous interrogeons *DBPedia* de deux manières sur les trois possibles (voir section 3.2.1)

3.1.1 IPTC

Du fait de leur nature hiérarchique et afin de formaliser l'ajout d'autres sources sémantiques aux liens plus complexes nous avons choisi de représenter les concepts sous la forme d'une table thématique et d'une

table association décrivant le type de relation ,la thématique étant associée à une clef étrangère décrivant le type de source. Nous séparons les concepts et thésaurus en quatre tables :

- *La thématique* : chargée de stocker le titre et la description du concept ainsi que de stocker l'ensemble des lemmes qui en sont issus.
- *Les relations* : qui peuvent se résumer à une table association.
- *Les thésaurus* : contenant les noms et descriptifs de la famille associée.
- *Les lemmes* : contenant les lemmes stockés un par un et associés à une thématique ou concept.

Cette représentation présente un triple intérêt. Nous conservons d'une part par une forme abstraite une modélisation souple applicable à différentes sources de données. Nous facilitons ensuite le travail de représentation sous forme de graphes explicités dans la section consacrée au prototype; enfin, nous facilitons le travail de pondération et de calcul de distance des mesures de similarité sémantique.

3.1.2 DBPedia

DBPedia permet trois types d'interrogation sur sa base. Ils mettent à disposition des « *dumps* » de leur base, un service par « *endpoint* » est accessible comme service Web et interrogeable par requête *SPARQL*, enfin des fichiers sous forme *XML* sont fournis pour une correspondance avec *Yago*. De ces méthodes d'extraction sus cités nous utilisons les deux dernières.
Du fait de la spécificité linguistique de notre projet nous nous intéressons aux sources disponibles dans la langue française. C'est pourquoi nous avons peuplé notre base automatiquement de quelques deux millions de concepts à partir d'un fichier XML référençant les ressources et au moyen d'une requête *SPARQL* filtrant sur les titres et les descriptions en français.
Nous interrogeons également l'« *endpoint* » de *DBPedia* pour établir une correspondance entre les termes que nous avons détectés comme ayant une forte probabilité d'être des noms propres et les concepts correspondants. (voir section 3.2).

3.1.3 Yago

Nous proposons d'utiliser l'ontologie *Yago* au travers d'un logiciel librement distribué en Java [12]. Pour ce faire nous utilisons à la fois le service de conversion des sources en SQL et le service de conversion en XML.

Le premier nous permet d'obtenir une table de lien de la forme {Concept}-{Relation}-{Concept} tel que {Einstein}-{a gagné}-{Prix Nobel} nous permettant par la suite d'utiliser les liens de relations dans notre mesure de similarité.

Le second nous permet d'obtenir un fichier de quelques deux millions de liens vers les ressources de *DBPedia* que nous injectons par la suite dans nos algorithmes de requête SPARQL pour en extraire une francisation des ressources sous la forme label et commentaires. Nous préconisons à ce sujet de ne pas utiliser la section résumé du fait de sa taille trop élevée qui atténuera la détection et la pondération des termes importants.
Ces deux utilisations nous permettent donc de faire le lien entre les ressources de *DBPedia* et de *Yago* tout en corrigeant le problème de la langue, « *Yago* » étant une ontologie anglophone et nos dépêches étant francophones.

3.2 Traitement des flux

L'actualité est par nature changeante, événementielle, ainsi sont donc nos flux. Toutefois nous proposons une série de traitements et de règles générales pouvant s'y appliquer et permettant le calcul de la proximité sémantique.
Nous procédons à l'agrégation de flux RSS dont la structure se compose d'un titre et d'une description que nous lemmatisons pour analyse et traitement. Ainsi on relèvera que les lemmes de titre sont par nature plus significatifs que les lemmes de description. (Le titre est une accroche qui représente l'essence de l'article). Un poids plus important doit donc être accordé aux lemmes associés aux titres par rapport aux lemmes associés à la description du contenu de l'article.
Basé sur l'article de J.Savoy [5] nous avons pu identifier un certain nombre de termes à la signification sémantique nulle (70 termes parmi lesquels des propositions, des noms, des verbes, etc.). On peut la rapprocher de la technique dite de discrimination dans le sens où ces termes n'apporteront pas de valeur ajoutée à notre algorithme de catégorisation. Nous proposons de filtrer et d'éliminer ces termes à la lemmatisation de nos articles.
Du fait même de la structure de nos thématiques statiques il semble important de prendre en compte le niveau hiérarchique. En effet, après analyse, nous observons une répétition entre les différents niveaux. Ainsi « M*usique Classique »* de niveau rappel dans son contenu les caractéristiques de son parent *Musique* de niveau deux. Ces deux spécialisations étant des dérivées de la super-catégorie *Art et spectacle* de niveau un et représentant des formes d'expression artistique. On retrouvera donc une occurrence de termes communs qui perdent de leur pertinence à la montée progressive du niveau d'abstraction.

Nous proposons donc une double approche sur la pondération des termes:

- *Déplier l'ensemble des sous concepts* : autrement dit, pour la détection du premier niveau, inclure dans notre recherche l'ensemble des lemmes des sous-ensembles pour le calcul de la probabilité de proximité sémantique.
- *Pondérer au fur et à mesure des itérations* : lorsque l'on « accroche » un nœud, réduire le poids des lemmes de la super-catégorie et augmenter ceux de la sous-catégorie. Cela nous permet à la fois d'obtenir une indication sur la justesse du chemin choisi par la conservation des propriétés acquises et de réduire l'influence des termes redondants.

Basé sur l'analyse manuelle des articles de presse [7] nous assignons comme prioritaire pour la détermination d'une catégorie de presse les noms propres et géographiques utilisés. En effet, un article ayant, par exemple, pour titre le nom d'un sportif ou un club de foot, verra ses probabilités d'être associé à la thématique « sport » beaucoup plus élevées.

Nous établissons donc une fonction de détection et de reconnaissance des noms propres couplés à l'ontologie *DBPedia* [4][8]. Pour ce faire nous proposons d'utiliser les propriétés d'analyse lexicale de *TreeTagger*. Les candidats à la catégorie que nous qualifions de nom propre ayant en commun la caractéristique de ne justement pas appartenir à un ensemble de catégorie lexicale. Après analyse, si un nom ou un adjectif est détectable comme tel, un nom propre n'a lui aucune des caractéristiques de ce dernier (nous reviendrons sur les résultats dans la section expérimentation).

Enfin en ce qui concerne la mesure de proximité sémantique nous proposons un approche mixe prenant en compte à la fois le poids des termes sus cité et les relations pouvant exister entre ces termes.

3.3 Calcul de proximité

L'ensemble de nos catégories sont représentés sous formes de graphes au sein de notre prototype. Cette classe abstraite contient une liste de ses nœuds parents et enfants ainsi qu'un ensemble de lemmes prétraités par *TreeTagger*. C'est sur cette représentation que notre système de calcul de proximité sémantique effectue une série d'opération permettant la détermination du nœud thématique le plus proche de l'article soumis.

[8] "DBPedia" http://dbpedia.org/About

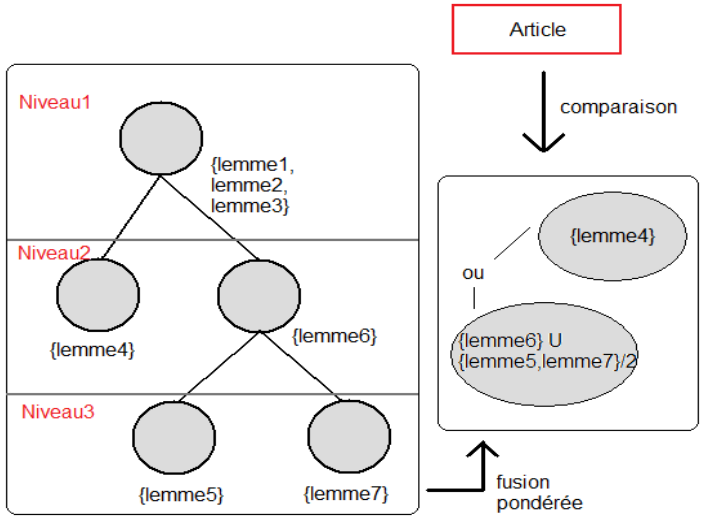

Figure 2 Processus de transformation par fusion pondérée des lemmes enfants

```
Entrée
          Niveaux à comparer
          Item RSS d'une dépêche
          L'ensemble des thématiques IPTC
Début
          Récupérer les lemmes de l'item RSS
          Pour chaque thématique de premier niveau
            récupérer les lemmes
            récupérer les nœuds enfants
            Si niveau >1 pour chaque nœud enfant
            récupérer les lemmes
              Si niveau > 2 pour chaque enfant des enfants
                  récupérer les lemmes.
          Comparer et retourner fréquence d'occurrence des lemmes de
          l'article   avec les lemmes des nœuds en divisant par deux
          pour chaque niveau d'éloignement du nœud à joindre.
Fin
```

Algorithme 1 Algorithme de comparaison

Pour ce faire notre algorithme récupère, des nœuds de plus haut degrés, les lemmes associés aux nœuds enfants et après avoir pondéré (on divise par deux la valeur pour chaque lemme de niveau inférieur), effectue un calcul de cooccurrence avec les lemmes de l'article. (Figure 2). De là il détermine le candidat le plus apte et répète l'opération sur les nœuds enfants.

L'originalité de l'algorithme tient principalement en sa procédure de « dépliage » des nœuds de niveaux inférieurs (nous pensons appliquer cette approche à des systèmes non hiérarchiques à l'avenir). Nous augmentons ainsi les chances de trouver des lemmes candidats à l'union tout en réduisant leur influence par pondération et par division de leur nombre.

4 Expérimentation et validation

Notre prototype écrit en Java dispose déjà de plusieurs fonctionnalités : l'agrégation de flux RSS d'un échantillon large, représentatif des différentes politiques de publication des sources de presse internet; la conversion et l'intégration de thésaurus; la lemmatisation et la pondération des termes liés à leur niveau hiérarchique à la fois pour les flux RSS et pour les thésaurus; l'identification des termes candidats au marquage en nom propre au moyen de la fonctionnalité d'analyse lexicale de *TreeTagger;* et enfin un algorithme de calcul de la proximité sémantique.

4.1 Agrégation

S'intéresser au pluralisme de l'information sur Internet implique d'intégrer les particularités de la publication de contenus sur le web, en comparaison des supports antérieurs comme l'imprimé ou l'audiovisuel. Car le processus de publication sur le web, ne se limite pas au modèle classique de diffusion mass-médiatique (sites de presse en ligne) : il comprend aussi le registre de l'auto publication (blogs), de la publication distribuée, et le niveau méta-éditorial (agrégateurs). Nous avons choisi 105 sources différentes représentatives de cette diversité:

- Des sites de presse en ligne : Le Monde, Libération, Le Figaro, etc.
- Des blogs : Plume de presse, Jean-Michel Apathie, etc.
- Des agences : Ria Novosti, etc.
- Des portails de news : Actualités Orange, MSN Actualités, etc.
- Des sites d'information participatifs : Rue89, etc.
- Des agrégateurs : Google Actualité, Wikio, etc.

Notre prototype propose au choix soit l'analyse des sorties du flux RSS d'une source sélectionnée, soit l'agrégation et la lemmatisation de l'ensemble. Nous convertissons le flux et l'insérons dans la base en en conservant la date, l'heure de publication (pour le regroupement

temporel), le titre, la description, et les termes extraits après lemmatisation.

4.2 Lemmatisation et filtrage

La lemmatisation se fait au travers d'un script utilisant *TreeTagger* qui permet accessoirement l'extraction linguistique ou le traitement automatique de la langue en convertissant notamment les verbes conjugués en leur forme infinitif. Notre prototype l'utilise pour trois de ses fonctionnalités, à savoir la lemmatisation des flux RSS, la lemmatisation des thésaurus (conversion linguistique) ainsi que la détection et le marquage des noms propres.

Comme explicité dans la section consacrée à notre contribution nous nous basons également sur le papier de J.Savoy [5] et éliminons soixante-trois lemmes vides de sens réduisant ainsi en moyenne le nombre de lemmes associé aux articles, définition et titre compris de 7%. Cette étape nous a permis de réduire légèrement les temps de calcul de proximité sémantique et d'alléger le poids de la base. Il est à noter que les lemmes en questions sont localisés à 80% dans les descriptions du fait de leurs natures plus verbeuses que les titres.

4.3 Proximités et résultats

Afin de tester la première étape d'implémentation de notre proposition de calcul de distance sémantique et d'aide à la catégorisation thématique des dépêches de presse nous avons sélectionné quarante articles aléatoirement sur un échantillon large de nos sources puis nous les avons catégorisés manuellement et avons récupéré en sortie les catégorisations proposées par le prototype selon que nous choisissions la fusion pondérée des nœuds enfants ou non.

Il est à noter que l'actualité récente (la grippe mexicaine) domine avec un tiers des dépêches orientés sur la santé.

Le Tableau 1présente les résultats obtenus. Les colonnes *sortie1, sortie2 et sortie3* représentent la catégorie retournée par l'algorithme de catégorisation en tenant compte des lemmes respectivement de niveau un, un et deux et des trois niveaux ensemble.

On observe dès les premières analyses que le niveau trois (à savoir prendre en compte toutes les lemmes de tous les descendants de chaque catégorie de niveau un et deux) donnent des résultats plus en adéquation avec nos attentes.

manuel	sortie1	sortie2	sortie3
Social	Police et justice	Police et justice	Police et justice
Santé	Alertes	Santé	Santé
Santé	Alertes	Santé	Santé
Politique	Santé	Santé	Santé
Santé	Police et justice	Police et justice	Police et justice
Santé	Alertes	Bulletins	Politique
Economie et finances	Société	Société	Economie et finances
Désastres et accidents	Politique	Politique	Désastres et accidents
Santé		Social	Politique
Politique	Police et justice	Police et justice	Politique
Société	Société	Société	Guerres et conflits
Sport	Sport	Sport	Sport
Police et justice	Politique	Politique	Sport
Police et justice	Police et justice	Police et justice	Police et justice
Santé	Social	Social	Politique
Santé	Politique	Politique	Sport
Police et justice	Politique	Politique	Sport
Politique	Politique	Politique	Politique
Economie et finances	rien	Social	Social
Economie et finances	Politique	Politique	Politique
Sport		Social	Sport
Santé	Santé	Santé	Santé
Politique	Politique	Politique	Politique
Politique	Politique	Politique	Politique
Social	Social	Social	Social
Police et justice	Politique	Politique	Sport
Sport	rien	Désastres et accidents	Sport
Santé	Désastres et accidents	Politique	Sport
Social	Vie quotidienne et loisirs	Vie quotidienne et loisirs	Social
Santé	rien	Santé	Santé
Guerres et conflits	Politique	Politique	Politique
Social	Vie quotidienne et loisirs	Vie quotidienne et loisirs	Social
Science et technologie	Gens animaux insolite	Gens animaux insolite	Social
Santé	Science et technologie	Politique	Politique
Environnement	Environnement	Police et justice	Environnement
Sport	Politique	Politique	Sport
Désastres et accidents	Désastres et accidents	Désastres et accidents	Désastres et accidents
Politique	Politique	Politique	Politique
Politique	Politique	Politique	Police et justice
Gens animaux insolite	Statistiques	Police et justice	Social

Tableau 1 Résultats attendus et résultats obtenus

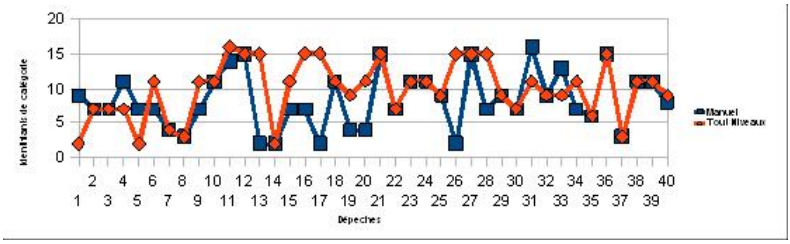

Figure 3 Correspondance avec la méthode de fusion pondérée

Une chose est à noter toutefois, notre catégorisation de référence est subjective. Bien que plus juste que celle obtenue par le prototype elle n'est pas exempte d'ambiguïté. En effet la frontière entre un mouvement social et une action politique est ténue et l'on trouvera dans l'article des références à ces deux thématiques.
Une dernière représentation booléenne de nos résultats (Figure 4 : somme des valeurs correctes en choisissant vrai pour thématique équivalente et faux dans le cas contraire divisé par le nombre total d'articles) vient toutefois confirmer notre analyse quant au taux de rappel entre la thématique attendue et les résultats de l'implémentation.

Methode	1	2	3
Satisfaction	30.00%	35.00%	55.00%

Figure 4 Taux de correspondance

Nous tirons trois conclusions de ces résultats :

- La prise en compte des nœuds enfants dans le choix des thématiques est une approche qui améliore significativement nos résultats.
- La quantité de lemmes contenus dans l'article influence la performance de l'appareillage. En effet la moitié des dépêches mal ou non catégorisées contenaient une description plus courte que la moyenne.
- Un apport sémantique est indispensable au bon fonctionnement de notre solution. Ce qui nous renforce dans notre idée d'enrichir le contenu des articles par des liens à *DBPedia*. Il semble clair que par la détection des concepts associés aux noms propres nous augmenterions considérablement la taille de l'ensemble de termes pertinents pour la catégorisation.

5 Conclusions et perspectives

Nous avons, au travers de ce papier, présenté notre approche d'aide à la catégorisation thématique des dépêches de presse francophones depuis des flux RSS, catégorisation basée sur un enrichissement sémantique permis d'une part par des sources externes (ontologies, thésaurus) et d'autre part par l'utilisation d'une mesure de calcul de proximité sémantique, mixte entre les méthodes par groupes de termes et les méthodes de calcul de chemin dans un arbre.

Les différents traitements successifs auxquels nous soumettons nos sources (lemmatisation, filtrage, analyse lexicale) nous permettent déjà d'obtenir un ensemble de propositions de rapprochement par catégorie thématique.

Nous projetons dans un prochain temps de fournir une réponse à la deuxième problématique de notre sujet à savoir de fournir une extension à la catégorisation temporelle des sujets. Au vu de l'état actuel de notre solution et des conclusions que nous avons tirées de nos recherches nous avons d'ores et déjà identifié trois facteurs de rapprochement, à savoir :
Le facteur thématique : On suppose que la probabilité de rapprochement d'articles à la thématique proche sera plus élevée que celle d'article à la thématique éloignée. Autrement dit la longueur du chemin qui sépare deux articles est un facteur déterminant de leurs proximités.
Le facteur temporel : du fait même de la nature événementielle des sujets de presse il apparaît essentiel dans une optique de catégorisation temporelle de déterminer un intervalle. De même la détection de ce dernier voit la probabilité du rapprochement de l'article à une famille dynamique de sujet augmenter. (Si l'on prend comme exemple le cadre des élections américaines, si l'on crée la section Obama [7] sous cette dernière les probabilités qu'un article contenant « *Obama* », « *américain* », « *élection* » ou n'importe quel lemme de ce type soient à rapprocher de cette catégorie s'en voient considérablement augmentées.)
Le facteur sémantique : La proximité et l'importance de certains lemmes utilisés pour la catégorisation thématique est un facteur de détermination et d'association au sujet. Dans le cadre même de regroupement au sein d'un nouveau sujet ces lemmes seraient choisis comme titre.

Nous projetons également une francisation et une normalisation finale des liens entre concepts intégrés à notre base, travail en cours qui devrait permettre de résoudre les problèmes de rapidité de traitement de correspondance actuel dû au fait du nombre élevé de tuples que représentent nos ontologies généralistes.

Enfin nous pensons ouvrir la voix à une implémentation d'une fonction d'apprentissage des termes importants relatifs aux évènements qui nous permettraient d'optimiser nos coûts en détection et en traitement.

6 Références bibliographiques

[1] H. Zargayouna. I*ndexation sémantique de documents XML,* Thèse section état de l'art, 2004

[2] M. Baziz, *Indexation Conceptuelle guidée par Ontologie pour la recherche d'Information*, Thèse pages 34-95, 2004

[3] E. Marty, F. Rebillard, N. Smyrnaios, A. Touboul, *Pluralisme et redondance ce l'information sur l'internet,* Revue Mot.

[4] R. Troncy, *Explorer des actualités multimédia dans le Web de Données*, IC 2009.

[5] J. Savoy, *Indexation et représentation comparative: Application au discours électoral*. CORIA 2009

[6] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, R. LeeMedia, *Meets Semantic Web - How the BBC uses DBpedia and Linked Data to make Connections*, 2009.

[7] A. Touboul, *Synthèse de l'étape « Analyse qualitative du traitement d'un sujet : Élection de Barack Obama» 2008.*

[8] Y. Matar, E. Egyed-Zsigmond, S. Lamji, *KWSim: Concepts Similarity Measure* CORIA 2008.

[9] A. Formica, *Concept similarity by evaluating information contents and feature vectors: a combined approach*. Commununications of the ACM 52 (2009) 145-149.

[10] F M. Suchanek, *Automated Construction and Growth of a Large Ontology* 2008.

[11] Fabian M. Suchanek, Gjergji Kasneci, GerhardWeikum, *YAGO:ALarge Ontology from Wikipedia andWordNet*, 2009.

[12] H. Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees* 2007.