# Analyse des données textuelles : Une approche d'extraction de contenu sémantique et un opérateur d'agrégation Top_KRankedTopics

Sarah Attaf*, Nadjia Benblidia*
Omar Boussaid **

*Laboratoire LRDSI Département d'informatique
Faculté des sciences université Saad Dahlab Blida,
route de Soumaa BP 270 Blida(09000)
sarah.attaf@gmail.com
benblidia@yahoo.com
** Laboratoire ERIC University of Lyon 2,
5 AV. P. Mends-France 69676 Bron Cedex Lyon, France
omar.boussaid@univ-lyon2.fr

**Résumé.** La prise en compte de la sémantique des données textuelles lors d'une analyse OLAP est une tâche complexe, qui n'est pas prise en charge par les systèmes décisionnels classiques. Pour répondre à cette problématique, nous proposons dans cet article une nouvelle approche pour l'extraction des descripteurs sémantique des données textuelles afin de les utiliser dans l'analyse. L'approche proposée est basée sur l'utilisation de la méthode Latent Dirichelet allocation (LDA) et la taxonomie Open Directory Project (ODP) comme une source de connaissance externe pour identifier les sujets pertinents dans un document textuel. Notre approche vise à construire pour chaque document textuel une hiérarchie sémantique à base des concepts du ODP. Pour prendre en compte cette hiérarchie sémantique lors d'une analyse OLAP, nous proposons une fonction de pondération ainsi qu'un opérateur d'agrégation qui sélectionne les $k$ premiers sujets et retourne pour chaque sujet une liste de documents pondérés.

## 1 Introduction

Le document électronique représente aujourd'hui un vecteur et un support d'information que les organisations ne doivent pas négliger. En effet, il est entendu que plus de 80 % des données nécessaires au fonctionnement d'une organisation sont encapsulées dans des documents, et non uniquement dans les bases de données opérationnelles. Ces données textuelles restent hors de portée des systèmes décisionnels, ce qui induit qu'une grande partie de l'information demeure inaccessible. Pour répondre à cette problématique et afin de pouvoir prendre profit des informations contenues dans ces documents, il est devenu plus que nécessaire d'intégrer ces données textuelles dans des systèmes d'information décisionnels permettant leur analyse.

Les systèmes décisionnels classiques ont déjà fait leurs preuves dans le domaine de l'analyse des données simples. Or ces systèmes ne sont pas adaptés à l'analyse des documents textes, ce qui met en évidence la nécessité de créer de nouveaux systèmes pour les données textuelles. L'entreposage de ces dernières demeure encore aujourd'hui une des difficultés majeures, et implique de nombreux problèmes, notamment ceux de leur modélisation et leur intégration d'une part et leur analyse d'autre part.

Les entrepôts de textes sont apparus comme une nouvelle solution, permettant une analyse multidimensionnelle des données textuelles. La nature complexe de ces données nécessite un traitement bien particulier, qui prend en compte leur sémantique. Dans la littérature, des méthodes de recherche d'information et de fouille de données ont donné de très bons résultats pour l'exploration des données textuelles. L'idée clef derrière ces entrepôts de textes est de faire un couplage entre les techniques de fouille de données et de recherche d'information d'un coté, et les techniques OLAP de l'autre coté. Ces derniers, permettent la navigation dans des cubes multidimensionnels d'une vue à une autre d'une manière interactive grâce aux opérateurs d'analyse de données OLAP qui accordent aux décideurs la possibilité d'exprimer des requêtes complexes et de donner une vision agrégée des résultats obtenus.

Dans la littérature, plusieurs travaux ont proposé des opérateurs d'agrégation pour agréger les données textuelles analysées, tel que TOP-KWK (Tournier et al., 2008) et le Tf-Idf adaptatif (Bringay et al., 2011). Cependant, la plupart de ces travaux sont basé sur le TF-IDF comme mesure d'analyse. Cela ne prend pas vraiment en compte la sémantique dans l'analyse OLAP des données textuelles. Par exemple, dans le cas des articles de presse, il est plus intéressant d'extraire des informations sur les sujets pertinents dans le document plutôt que d'extraire les termes les plus fréquents. Dans cet article, nous proposons : (1) une approche pour l'extraction des descripteurs sémantique dans un document textuel, (2) un opérateur d'agrégation pour l'analyse des données textuelles( Top_KRankedTopics), qui sélectionne les $k$ premiers sujets et retourne pour chaque sujet une liste de documents pondérés.
La suite de l'article est organisée comme suit. La section 2 présente un état de l'art des travaux portant sur l'analyse multidimensionnelle des données textuelles ainsi qu'une étude comparative entre ces travaux. La section 3 présente notre approche d'extraction de sujets pertinents dans un documents textuel. La section 4 expose notre fonction de pondération qui permet de pondérer les sujets (topics) en fonction de leur représentativité dans les documents à agréger. La section 5 définit la fonction d'agrégation Top_KRankedTopics. Enfin la section 6 conclut l'article.

# 2 État de l'art

Dans les dernières années, le domaine de l'entreposage et l'analyse en ligne OLAP a connu un grand nombre de travaux traitants les données complexes. Ces travaux couvrent les différents aspect de stockage et d'analyse ; nous citons les travaux sur : l'intégration des données web( Bhowmick et al. (2003), Xyleme (2001)) ; les entrepôt de données multimédia (Pissaloux et al. (2001), Arigon et al. (2007), Vanea et Potolea (2011), Bleyberg (2000), McCabe et al. (2000)) ; les données semi-structurées représentés en XML(Extensible Markup Language)(Golfarelli et al. (2001), Vrdoljak et al. (2003), Park et al. (2005)) ; et le stockage des

données non structurées (Inokuchi et Takeda (2007) ,Keith et al. (2006))...etc.
Les données textuelles représentent un type particulier des données complexes. Afin de les considérer dans l'analyse multidimensionnelles, plusieurs travaux ont été élaborés. Ces travaux sont groupés selon Attaf et Benblidia (2013) en deux catégorie :

1. Travaux avec des modèles extensifs, qui proposent d'étendre les modèles d'entrepôts traditionnels pour permettre l'analyse des données textuelles. Parmi ces travaux nous trouvons ceux qui proposent d'étendre les modèles d'entrepôts classiques en intégrant une dimension sémantique, tel que :

   - une hiérarchie de sujets dans *Topic Cube* Zhang et al. (2009), qui définit la hiérarchie de sujets '*Topics*' comme étant une dimension d'analyse et propose deux mesures probabilistes : la distribution d'un mot dans un thème *word distribution of a topic* $p(wi)$ et la couverture d'un thème par les documents *topic coverage by documents* $p(topic.j)$. La couverture d'un *topic* est la probabilité qu'un document $dj$ couvre le *topic*. Ainsi, nous pouvons facilement prédire quel est le sujet dominant dans l'ensemble des documents en agrégeant la couverture sur tous les documents dans l'ensemble.
   - une hiérarchie de termes dans *Text Cube*, qui spécifie les relations sémantiques entre les termes textuels extraits des documents, ce qui permet une navigation sémantique dans les données textuelles grâce aux deux opérateurs qui lui sont associés : *pull-up* and *push-down*.
   - une *AP-Structure* basée sur les items fréquent nommée *AP-Sets* dans Bautista et al. (2010), obtenus par l'application de l'algorithme *apriori* sur les attributs textuels d'une base de données transactionnelle.

   D'autre travaux ont basé leurs modèles sur la technique de classification tel que :(i) *Doc Cube* Mothe et al. (2003), qui permet de produire des vues globales de grands corpus de documents, en utilisant la classification. Son élément de base est l'utilisation du concept hiérarchie afin de structurer les collections de documents, chaque hiérarchie correspond à une facette de documents 'dimension d'analyse' pour laquelle les utilisateurs peuvent être intéressés.(ii) *microtextcluster* Zhang et al. (2011), qui propose proposé d'introduire une nouvelle mesure d'analyse $(mean_i, size_i)$ qui représente respectivement le vecteur *mean* (un vecteur de termes pondérés) et la taille d'un *micro-cluster*, où un *micro-cluster* est une cellule texte qui permet de compresser les documents similaires (chaque cellule texte contient un certain nombre de documents). Cette compression (en *micro-cluster*) permet de retenir des informations sémantiques essentielles sur les cellules textuelles.

2. Travaux avec des modèles basés sur de nouveaux concepts. Parmi ces travaux nous citons le modèle en galaxie Tounier (2007), le modèle d'objets complexes Boukraa et al. (2011) et le modèle multidimensionnel sémantique d'objets textes (MSMTO) Attaf et al. (2014).
   Le modèle en galaxie est basé sur le concept galaxie. Une galaxie est définie comme étant un regroupement de dimensions liées entre elles par un ou plusieurs noeuds centraux ; chaque noeud modélise les dimensions compatibles pour une même analyse. Con modèle est basé sur la généralisation du concept de constellation de Kimball (Kimball, 1996).

Cette approche consiste à décrire un schéma multidimensionnel par l'unique concept de dimension où la notion de fait est supprimée. Le modèle d'objets complexe est basé sur le paradigme objet grâce auquel il est possible de représenter les objets de l'univers et de capter la sémantique qu'ils véhiculent, notamment dans les liens avec les autres objets. Ainsi ils modélisent le monde réel par un ensemble d'objets complexes qui décrivent les entités de ce dernier. Le modèle MSMTO intègre un nouveau concept contenu sémantique (*semantic content object*) qui permet de représenter la sémantique des données textuelles et de l'organiser sous forme hiérarchique, pour assurer une analyse sémantique sur différents niveaux de granularité.

Nous présentons dans ce qui suit, une étude comparative sur les différents travaux cités auparavant. Nous comparons l'ensemble de ces travaux par rapport à la prise en compte des cinq aspects suivants :

a. **L'aspect structurel :** la modélisation des données textuelles dans un but d'analyse peut considérer le document texte comme étant une donnée élémentaire. L'objectif consiste alors de structurer et de stocker les documents dans une base de documents textes et de les préparer à l'analyse, sans prendre en compte la structure interne des documents. Toutefois, cette approche de modélisation ne répond pas à toutes les exigences d'un décideur, tel que l'analyse des sections sportives d'un ensemble de journaux. Ce type d'analyse n'est pas supporté par cette approche car la structure interne des documents qui divise le document en plusieurs niveaux hiérarchiques, ce qui permet une analyse sur de différents niveaux de granularité, n'est pas prise en considération. Ainsi nous définissons un modèle qui prend en compte l'aspect structurel des documents, et permettant une analyse multidimensionnelle sur de différents niveaux structurels.

b. **L'aspect sémantique :** l'extraction et la représentation de la sémantique véhiculée dans les données textuelles présentent une problématique déjà traitée dans la littérature dans les domaines d'extraction de connaissances et de la recherche d'information. Tandis que dans les entrepôts de données, la prise en compte de cet aspect important dans la modélisation multidimensionnelle est une nouvelle problématique. Répondre à cette problématique reviens à trouver une manière d'incorporer la sémantique des données textuelles et de la modéliser au sein d'un cube de données.

c. **La flexibilité d'analyse :** dans les systèmes décisionnels classiques un fait représente un sujet d'analyse prédéfini. La définition d'un fait rend la spécification d'analyses peu flexible, car le décideur se voit contraint d'employer ces faits comme sujets. La flexibilité d'analyse est apparue comme un nouveau besoin exprimé par les décideurs. Elle réside dans le fait où le sujet d'analyse n'est pas prédéfini au préalable mais choisi au moment de l'analyse. Dans le domaine de l'analyse des données textuelles, nous percevons que le problème de flexibilité est assez complexe. Ainsi nous posons cette problématique autrement, lors d'une analyse textuelle, le contenu sémantique de ces données peut être vu comme étant une mesure d'analyse *(K-top keyword,Topic)*. Comme il peut être considéré comme étant un axe d'analyse. Donc assurer une bonne flexibilité revient à donner à ce contenu sémantique un double rôle.

d. **Mesure textuelle** : la modélisation reposant sur les concepts de fait et de dimension associés à des indicateurs numériques permet des analyses simples de documents textes.

Ces analyses reposent principalement sur le comptage de documents. Une bonne analyse de contenu des données textuelles doit prendre en compte les mesures textuelles.

e. **Opérateur OLAP spécifiques aux données textuelles** : les opérateurs OLAP appliqués aux données simples ne sont pas adaptés aux données textuelles. Les fonctions d'agrégation numériques telles que $somme$, $moyenne$ s'appliquent bien sur des données numériques, mais ne permettant pas d'agréger les données textuelles. Donc définir de nouveaux opérateurs OLAP s'appliquant sur les données textuelles s'avère nécessaire.

Le tableau ci-dessous, présente une étude comparative des modèles présentés ci dessus.

| Modèles d'entrepôts de textes | Familles de modèles | | Mesure texte | Opérateurs OLAP | | Aspect Sémantique | Aspect structurel | Flexibilité d'analyse |
|---|---|---|---|---|---|---|---|---|
| | Modèles extensifs | Modèles à nouveaux concepts | | Fonctions d'agrégation | Opérateurs de navigation | | | |
| *E.documents* Khrouf et Dupuy (2001) | | X | - | - | - | - | X | Bonne flexibilité |
| *Mire* Lee et al.(2002) | X | | X | - | Drill down et Roll-up | - | - | Non flexible |
| *DocCube* Lee et al.(2002) | X | | - | Score(Dd) | Drill down et Roll-up | X | - | Non flexible |
| *D.cube* Tseng et al.(2006) | X | | - | Count | Drill down et Roll-up | - | - | Non flexible |
| *Galaxie* Tounier(2007) | | X | X | AVG-KW, Top-KW | Drill down et Roll-up | X | X | Bonne flexibilité |
| *TextCube* Lin et al.(2008) | X | | X | - | pull-up et push-down | X | - | Non flexible |
| *TopicCube* Zhang et al.(2009) | X | | X | - | Drill down et Roll-up | X | - | Non flexible |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *MMAP-structure* Bautista et al.(2010) | X | | - | - | - | X | - | Non flexible |
| *M.Cube* Zhang et al.(2011) | X | | - | - | - | X | - | Non flexible |
| *MMOC* Boukraa et al.(2011) | | X | X | utilisation du Top-KW | Drill down et Roll-up | - | X | Bonne flexibilité |
| *MSMTO* Attaf et al.(2014) | | X | X | utilisation du Top-KW | Drill down et Roll-up | - | X | Bonne flexibilité |

TAB. 1: TABLEAU COMPARATIF

Malgré que ces travaux ont permis d'effectuer des analyses multidimensionnelles sur les données textuelles, nous constatons qu'ils sont toujours limités et ne traitent que quelques aspects de complexité liés à l'analyse de ce type de données.
La prise en compte de l'aspect sémantique des données textuelles lors d'une analyse OLAP est devenu primordial. Toute fois, l'extraction de ce contenu sémantique demeure toujours une des difficultés majeur qui implique de nombreux défis. Afin de traiter cette problématique nous proposons dans cet article une nouvelle méthode pour l'extraction des sujets pertinents dans un document textuel. l'approche proposée est basé sur l'utilisation de la méthode Latent Dirichelet allocation (LDA)Blei et Jordan (2003) et la taxonomie Open Directory Project (ODP) comme une source de connaissance externe pour identifier les sujets pertinents dans un document texte. Nous proposons aussi un opérateur d'agrégation Top_KRankedTopics qui restitue au décideur les $K$ sujets les plus pertinents d'un ensemble de sujets à agréger.

# 3 Approche d'extraction du contenu sémantique des documents texte

Dans cette section nous présentons notre approche d'extraction du contenu sémantique des données textuelles.
Nous définissons le contenu sémantique comme étant l'ensemble des sujets pertinents, des phrases pertinentes,...etc. Pour détecter ce contenu sémantique de nombreuses applications proposent d'utiliser des modèles de sujets(topic models), qui sont une suite d'algorithmes permettant de dévoiler la structure thématique caché dans des collections de documents. Ces algorithmes nous aident à développer de nouvelles façons de rechercher, parcourir et résumer les grandes archives de textes. Une variété de modèles probabilistes de sujet comme LDA (Latent Dirichelet allocation) Blei et Jordan (2003), PLSA (Probabilistic latent semantic allocation) Zhang et al. (1999) ou LSA (Latent semantic analysis) Landauer et Dumais (1997) ont été utilisées pour analyser le contenu des documents et le sens des mots. Ces modèles

utilisent la même idée fondamentale, qu'un document est un mélange de sujets. Chaque sujet est représenté par une distribution de mots $T_i(word_j/P(word_j|T_i)); i, j \in \mathbb{N}$ et la probabilité $P(T_i|Doc_k); i, k \in \mathbb{N}$, tel que $P(word_j|T_i)$ est la probabilité du mot $word_j$ dans le sujet $Topic_i$ et $P(T_j|Doc_k)$ est la probabilité du sujet $Topic_j$ dans le document $Doc_k$. Toute fois, la distribution de mots donnée pour chaque sujet nous nous permet pas d'extraire les concepts sémantique des sujets. Pour traiter cet aspect, nous proposons une approche qui associe l'utilisation de la méthode LDA avec la taxonomie ODP afin de reconstruire la sémantique de la distribution de mots. L'approche proposée est divisée en trois phases principales présentées dans la figure 1.

## 3.1 La préparation de données

Dans cette phase, un ensemble de traitements sont enchainés sur le texte du document, afin de filtrer les termes pertinents en éliminant les segments de texte (tokens) non pertinents. Les principales tâches effectuées concernent :

1. La Tokenisation ou segmentation du texte ; elle consiste à parcourir le texte en vue de récupérer les termes et supprimer les caractères spéciaux et la ponctuation.

2. L'élimination des mots vides ; consiste à éliminer les mots outils de la langue en utilisant une liste de mots vides.

3. La Racinisation ou stemming ; consiste à trouver la racine des verbes fléchis et à ramener les mots pluriels et/ou féminins au masculin singulier.

**Exemple.** Considérons le document $doc_1$ composé du texte suivant : "ETL is a process in data warehousing responsible for pulling data out of the source systems and placing it into a data warehouse." . En appliquant la phase préparation de données, on obtiendra : $doc_1$= "be process data warehouse responsible pull data source system place data warehouse"

## 3.2 Extraction des domaine ODP des documents texte

Dans cette étape, nous identifions pour chaque document texte, son domaine ODP en utilisant textwise API [1]

**Exemple.** Considérons les documents $Doc_1; Doc_2; ...Doc_n$. Par identifier les domaines des documents nous obtenons $ODP\_Domain(Doc_1 = Sport); ODP\_Domain(Doct_2 = Society)...ODP\_Domain(Doc_n = ....)$..etc.

## 3.3 Extraction des sujets via LDA

Afin d'extraire les sujets cachés dans le corpus de texte, nous appliquant le modèle LDA sur les document nettoyé. Pour ce fait, nous devons définir le nombre de sujets et le nombre de mots attribués à chaque sujet. Alors, chaque sujet est représenté par une distribution de mots, ex : $T_0\{low(0.5); stadium(0.8); sport(0.7); major(0.2); fight(0.2); football(0.4)\}$ avec une matrice qui définit le poids de chaque sujet dans chaque document :

[1] http ://www.textwise.com/demo

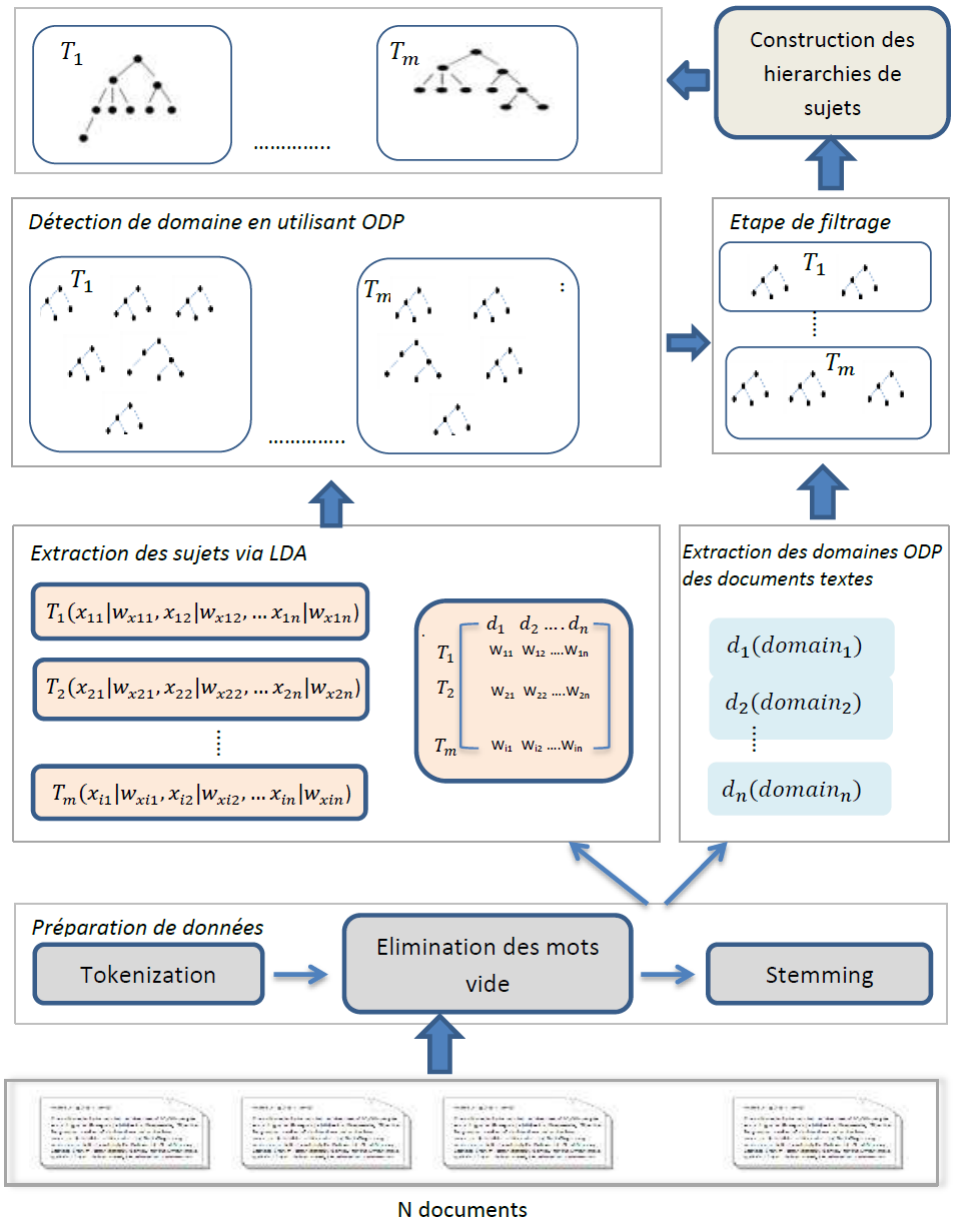


FIG. 1 – *System Architecture*

## 3.4 Détection de domaine en utilisant ODP

Les sujets obtenus en appliquant la méthode LDA sont donnés par une distribution de mots, alors on ne peut pas identifier les concepts sémantique de ses sujets. Pour répondre

$$\begin{array}{c} \\ T_1 \\ T_2 \\ ... \\ T_m \end{array} \begin{array}{c} \begin{array}{cccc} Doc_1 & Doc_2 & ... & Doc_k \end{array} \\ \begin{pmatrix} 0,6 & 0,1 & ... & 0,3 \\ 0,2 & 0,6 & ... & 0,2 \\ 0,35 & 0,05 & ... & 0,6 \\ 0,35 & 0,05 & ... & 0,6 \end{pmatrix} \end{array}$$

TAB. 2 – *Matrice représentant le poids de chaque sujet par document*

à cette problématique, nous proposons de présenter chaque sujet comme une distribution de domaines, en identifiant pour chaque mot dans la distribution d'un sujet tous les domaine ODP avec leurs différents niveau. Afin de présenter chaque sujet par un ensemble de hiérarchies de domaines basé sur ODP, nous générons pour chaque mot toutes les hiérarchies possible à partir de la taxonomie ODP.

## 3.5 Etape de filtrage

Un mot représente une unité très spécifique et peut être connecté à plusieurs domaine ODP. Alors, certaines hiérarchies de concepts obtenues en appliquant l'étape précédente, peuvent être considéré comme non pertinentes pour un document. Comme les domaines de document ainsi que les hiérarchies de concepts représentants les sujets sont basés sur ODP, nous définissons une hiérarchie de concepts comme étant non pertinente pour un document donné, si la racine de cette dernière est différente du domaine du documents. Dans cette étape, nous éliminons toutes les hiérarchies non pertinente.

**Example.** Considérons le document $doc_1$ avec domaine égale à $Sport$ et le sujet $Topic_0$ donné par la distribution de mot suivante : $Topic_0$\{low(0.5) ; stadium(0.8) ; sport(0.7) ; major(0.2) ; fight(0.2) ; football (0.4)\}. La figure 2 présente un exemple de hiérarchies pertinentes et non pertinente basé sur ODP pour le mot *Stadium*.

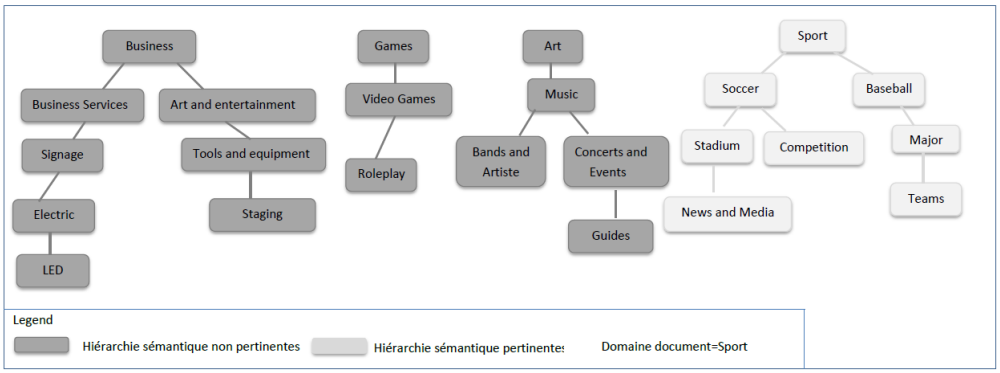


FIG. 2 – *Exemple de hiérarchies pertinentes et non pertinentes pour le mot Stadium*

### 3.6 Construction des hiérarchies sémantiques

Dans cette phase, nous construisons pour chaque document textuel une hiérarchie sémantique, en combinant les hiérarchies de concepts qui lui ont été identifier par l'étape de filtrage.

## 4 Fonction de pondération

Dans la section précédente, nous avons proposé une approche pour la construction d'une hiérarchies sémantique basé sur les concepts de la taxonomie ODP, pour donner une représentation sémantique de chaque document textuel. Dans cette section, nous présentons une fonction de pondération qui calcule la probabilité de chaque concept dans chaque document texte. Notre fonction de pondération est décrit par la formule suivante :

$$P(Concept_j, Document_m) = \sum_{k=1}^{n} \sum_{i=1}^{l} \frac{Nconcept_j}{Nconcept} * P(w_i|topic_k) * P(topic_k|Doc_m) \quad (1)$$

tel que :

- $Nconcept_j$ est le nombre d'occurrence du $concept_j$ avec le mot $word_i$ en utilisant la taxonomie ODP.
- $Nconcept$ est le nombre de tous les concepts avec le mot $word_i$ en utilisant ODP.
- $P(w_i|topic_k)$ est la probabilité du mot $word_i$ dans $topic_k$.
- $P(topic_k|Doc_m)$ est la probabilité du sujet $topic_k$ dans le document $Doc_m$.

## 5 Opérateur d'agrégation $Top\_KRankedTopics$

Notre opérateur d'agrégation $Top_K RankedTopics$ agrège un ensemble de document par : (1) sélectionner les K sujets les plus pertinent (ayant le plus grand poids), (2) pondérer ces documents selon la hiérarchie sémantique qui les représente(obtenue en appliquant l'approche décrit dans la section 3). Notre fonction d'agrégation est basé sur la fonction de pondération donnée par (1) pour le calcul du poids. Notre opérateur d'agrégation sélectionne les $k$ premiers sujets et retourne pour chaque sujet une liste de documents pondérés.

**Exemple.** Considérons une requête $q_1$, dans laquelle le décideur veut analyser les deux sujet les plus pertinent dans les journaux de presse pour date= {"2014" ;"2013"} ; Location={"Syria" ;"France"}. les résultats présentés dans la figure 3 sont obtenus par l'application de notre opérateur d'agrégation.

## 6 Expérimentation

Pour valider nos propositions, une étape d'évaluation est plus que nécessaire. la section 6.1 présente le corpus de données utilisé dans nos expérimentation. Les sections 6.2 décrit notre protocole expérimental.

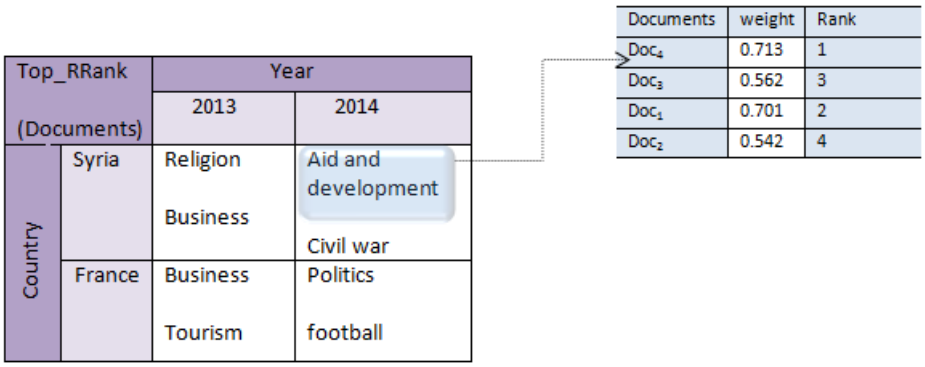


FIG. 3 – *Exemple d'analyse utilisant Top_KRankedTopics*

TAB. 3 – *corpus de données*

| Mois | Totale des documents | Totale des mots |
|---|---|---|
| 1 | 254 | 76300 |
| 2 | 595 | 105450 |
| 3 | 510 | 89502 |
| 4 | 300 | 82404 |
| 5 | 405 | 81318 |
| 6 | 577 | 102005 |

## 6.1 Présentation du corpus

Pour faire nos expérimentations, nous avons constitué un corpus de taille moyenne d'article de presse. Nous avons utilisé pour cela le site Europe Topics sur le premier semestre de l'année 2014 [1]. Nous avons utilisé apache[1] pour extraire le texte des fichiers html. Nous avons séparé les données sur 6 mois, comme indiqué sur le tableau 3.

## 6.2 Protocole expérimental

Notre protocole expérimentale peut être divisé on trois principaux procédures :

- Nous construisons un cube de données sémantique basé sur notre modèle MSMTO.
- Nous implémentons notre méthode d'extraction de contenu sémantique afin d'alimenter la dimension sémantique dans notre cube de données.
- Nous effectuons un teste pour évaluer notre opérateur d'agrégation

[1]http ://www.eurotopics.net/fr/home/presseschau/aktuell.html
[1]Ti-ka23http ://www.tika.apache.org

## 6.3 Résultats

Afin d'évaluer notre opérateur d'agrégation $Top\_KRankedTopics$, nous avons mené une série de testes sur un cube de données construit en se basant sur le modèle MSMTO. Nous avons demandé à un décideur d'émettre des requêtes d'analyse. Nous considérons une requête, dans laquelle, le décideur veut analyser les deux sujets pertinents des article de presse pour : Time = "2014", Country = "Algeria". Les résultats sont présentés dans la figure 3.

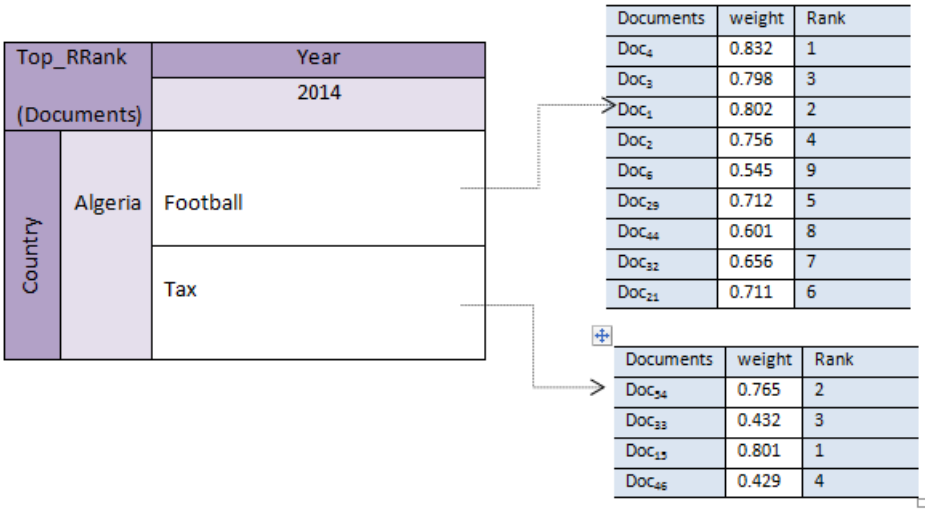


FIG. 4 – *Évaluation du Top_KRankedTopics*

Après avoir examiner les résultats, le décideur juge qu'ils sont pertinent à sa requête d'analyse.

# 7 Conclusion

Analyser les données textuelles afin de pouvoir tirer profit des informations qu'elles contiennent est devenu essentiel, à cause de leurs volumes importants et de la quantité d'information qu'elles contiennent. Nous avons présenté dans cet article un état de l'art ainsi qu'une étude comparative sur différents travaux traitants l'analyse des données textuelles. Nous avons proposé aussi une nouvelle approche pour l'extraction du contenu sémantique des documents texte. L'approche proposée vise à construire pour chaque document texte une hiérarchie sémantique qui le représente. Afin de mieux profiter de cette hiérarchie sémantique, nous avons proposé une fonction de pondération et un opérateur d'agrégation qui sélectionne les $k$ premiers sujet et retourne pour chaque sujet une liste de documents pondérés.

# Références

Arigon, A., M. Miquel, et A. Tchounikine (2007). Multimedia data warehouses : a multiversion model and a medical application. *Multimedia Tools Appl. 35*(1), 91–108.

Attaf, S. et N. Benblidia (2013). Modelisation multidimensionnelle des donnees textuelles ou en sommes-nous ? In *ASD Conference Proceedings*, pp. 3–25. Conference maghrebine sur les avancees des systemes decisionnels.

Attaf, S., N. Benblidia, et O. Boussaid (2014). The multidimensiional semantic model of text objects : A frame work for text data analysis. In *Lecture Notes in Computer Science (LNCS)*, pp. 3–25. Internationnal Conference on Model ans Data Engineering.

Bautista, M., C. Molina, E. Tejeda3, et A. Vila (2010). Using textual dimensions data warehousing processes. In *International Conference, IPMU, Dortmund, Germany*, pp. 158–167. IPMU.

Bhowmick, S. S., S. K. Madria, et W. K. Ng (2003). Representation of web data in A web warehouse. *Comput. J. 46*(3), 229–262.

Blei, D.M.and Ng, A. et M. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research 3*(2), 993–1022.

Bleyberg, M., a. G. K. (2000). Dynamic multi-dimensional models for text warehouses in : Systems, man, and cybernetics. In *Lecture Notes in Computer Science (LNCS)*, pp. 2045–2050. IEEE International Conference.

Boukraa, D., O. Boussaid, F. Bentayeb, et D. Zegour (2011). Modle multidimensionnel d'objets complexes : Du modele d'objets aux cubes d'objets complexes. *Ingénierie des Systèmes d'Information 16*.

Bringay, S., N. Béchet, F. Bouillot, P. Poncelet, M. Roche, et M. Teisseire (2011). Analyse de gazouillis en ligne. In *Actes des 7èmes journées francophones sur les Entrepôts de Données et l'Analyse en ligne, Clermont-Ferrand, France, EDA 2011, Juin 2011*, pp. 87–102.

Golfarelli, M., S. Rizzi, et B. Vrdoljak (2001). Data warehouse design from XML sources. In *Proceedings of the 4th ACM International Workshop on Data Warehousing and OLAP (DOLAP 2001), Atlanta, Georgia, USA, November 9, 2001*, pp. 40–47.

Inokuchi, A. et K. Takeda (2007). A method for online analytical processing of text data. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6-10, 2007*, pp. 455–464.

Keith, S., O. Kaser, et D. Lemire (2006). Analyzing large collections of electronic text using OLAP. *CoRR abs/cs/0605127*.

Kimball, R. (1996). *The data warehouse toolkit: Practical Techniques for Building Dimensional Data Warehouses,*. John Wiley and Sons.

Landauer, T. et S. Dumais (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review 104*, 211 – 240.

McCabe, M. C., J. Lee, A. Chowdhury, D. A. Grossman, et O. Frieder (2000). On the design and evaluation of a multi-dimensional approach to information retrieval. In *SIGIR*, pp. 363–365.

Mothe, J., B. Chrisment, C.and Dousset, et J. Alaux (2003). Doccube: Multi-dimensional visualisation and exploration of large document sets. *Journal of the American Society for Information Science and Technology 54*, 650–659.

Park, B., H. Han, et I. Song (2005). XML-OLAP: A multidimensional analysis framework for XML warehouses. In *Data Warehousing and Knowledge Discovery, 7th International Conference, DaWaK 2005, Copenhagen, Denmark, August 22-26, 2005, Proceedings*, pp. 32–42.

Pissaloux, E., J. You, J. Liu, et T. S. Dillon (2001). On hierarchical multimedia information retrieval. In *ICIP (2)*, pp. 729–732.

Tounier, R. (2007). *Analyse en ligne (OLAP) de documents*. Thèse de doctorat, Universit Toulouse III . Paul Sabatier.

Tournier, R., vGeneviève Pujolle, F. Ravat, et O. Teste (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.

Vanea, A. et R. Potolea (2011). Semantically enhancing multimedia data warehouses - using ontologies as part of the metadata. In *ICEIS 2011 - Proceedings of the 13th International Conference on Enterprise Information Systems, Volume 1, Beijing, China, 8-11 June, 2011*, pp. 163–168.

Vrdoljak, B., M. Banek, et S. Rizzi (2003). Designing web warehouses from XML schemas. In *Data Warehousing and Knowledge Discovery, 5th International Conference, DaWaK 2003, Prague, Czech Republic, September 3-5,2003, Proceedings*, pp. 89–98.

Xyleme, L. (2001). A dynamic warehouse for XML data of the web. *IEEE Data Eng. Bull. 24*(2), 40–47.

Zhang, D., C. Zhai, et J. Han (1999). Probabilistic latent semantic indexing. pp. 50–57. SIGIR.

Zhang, D., C. Zhai, et J. Han (2009). Topic cube: Topic modeling for olap on multidimensional text databases. In *SDM '09: Proceedings of the 2009 SIAM International Conference on Data Mining, Sparks, NV, USA"*, pp. 1124–1135. SDM 09.

Zhang, D., C. Zhai, et J. Han (2011). Mitexcube:microtextcluster cube for online analysis of text cells. pp. 204–218. The NASA Conference on Intelligent Data Understanding (CIDU).

# Summary

The consideration of textual data semantic in OLAP analysis is a complex task, which is not supported by traditional business intelegence systems. To address this problem, we propose a new approach for semantic descriptors extraction of textual data for analysis purpuses. The proposed approach is based on the use of Latent Dirichelet allocation method (LDA) and Open Directory Project (ODP) taxonomy as an external source of knowledge, to identify relevant topics in text documents. Our approach is to build for each text document a semantic hierarchy based on ODP concepts. To make this semantic hierarchy usfull in an OLAP analysis; we propose a weighting function and an aggregation operator that selects the first $k$ subject and returns for each