Indexation automatique des textes arabes : état de l'art

Automatic indexing of Arabic documents: State of the art

Mohamed Salim El Bazzi
Laboratoire IRF-SIC, Université Ibn Zohr, Agadir, Maroc
elbazzi.mohamedsalim@edu.uiz.ac.ma

Taher Zaki
Laboratoire IRF-SIC, Université Ibn Zohr, Agadir, Maroc
t.zaki@uiz.ac.ma

Driss Mammass
Laboratoire IRF-SIC, Université Ibn Zohr, Agadir, Maroc
mammass@uiz.ac.ma

Abdelatif Ennaji
Laboratoire LITIS, Université de Rouen, Rouen, France
abdel.ennaji@univ-rouen.fr

Résumé

L'indexation des documents est une phase cruciale dans le processus de fouille de textes. Elle permet de représenter les documents par les descripteurs les plus pertinents vis-à-vis de leurs contenus. À ce propos, plusieurs approches sont proposées dans la littérature, notamment pour l'anglais, mais elles sont inexploitables par les documents en langue arabe en raison de ses caractéristiques spécifiques, de sa richesse morphologique et grammaticale et de son vocabulaire. Cet article dresse un état de l'art des méthodes d'indexation et de leurs apports à la langue arabe. Nous proposons une catégorisation des travaux selon les approches et les méthodes les plus utilisées en indexation automatique de documents textuels. Nous avons adopté une sélection qualitative des articles. Ainsi, avons-nous retenu les travaux constituant des contributions significatives au niveau de l'indexation et présentant des résultats considérables.

Abstract

Document indexing is a crucial step in the text mining process. It is used to represent documents by the most relevant descriptors of their contents. Several approaches are proposed in the literature, particularly for English, but they are unusable for Arabic documents, considering its specific characteristics and its morphological complexity, grammar and vocabulary. In this paper, we present a reading in the state of the art of indexation methods and their contribution to improve Arabic document's processing. We also propose a categorization of works according to the most used approaches and methods for indexing textual documents. We adopted a qualitative selection of papers and we retained papers approving notable indexation contributions and illustrating significant results.

Mots-clés

Fouille de textes, langue arabe, sémantique, méthode d'indexation, classification.

Keywords

Text mining, Arabic language, semantic, indexation methods, classification.

1. Introduction

La masse documentaire disponible sur internet et la numérisation des documents textuels ne cessent d'augmenter. Ce changement révolutionnaire présente de grands défis, et en même temps de grandes opportunités aux chercheurs pour exploiter les informations cachées en introduisant différentes approches.
La plupart des travaux effectués dans ce domaine ont été consacrés surtout aux langues occidentales, notamment l'anglais. En revanche, la langue arabe, étant une langue riche morphologiquement et fortement flexionnelle, a connu peu d'études au niveau de l'extraction des descripteurs. Ceci est dû au problème majeur de la complexité de son traitement automatique.
L'indexation des documents consiste à extraire les mots-clés qui représentent le mieux un document. Malgré le rôle primordial de cette phase dans la suite du processus de fouille et d'analyse des textes, peu sont les travaux recensés à ce niveau (Zaki, 2013). Cet article présente une lecture dans l'état de l'art des différentes méthodes d'extraction des descripteurs, ainsi que leurs applications et leurs compatibilités avec la langue arabe.
Le reste de cet article est organisé comme suit. La deuxième partie introduit le processus d'indexation. La troisième partie est dédiée à la présentation des différentes approches de sélection de descripteurs. Nous entamons une discussion dans la quatrième partie et enfin nous concluons par l'apport de ces approches au développement du traitement automatique de la langue arabe.

2. L'indexation des documents textuels

Indexer un document revient à élire ses descripteurs les plus représentatifs afin de générer la liste des termes d'indexation. C'est un moyen de retrouver l'ensemble des termes caractérisant un document. L'indexation des documents est une étape primordiale dans le processus de fouille de textes car elle détermine de quelle manière les connaissances contenues dans les documents sont représentées (Zaki, 2013) (Mountassir, 2012). Elle a lieu à chaque ajout d'un document dans le corpus étudié.
L'AFNOR définit l'indexation aussi comme le «Processus destiné à décrire et à caractériser, au moyen des termes ou indices d'un langage documentaire ou au moyen des éléments d'un langage naturel (libre), des données résultant de l'analyse du contenu d'un document (ressource, collection) ou d'une question, en vue d'en faciliter la recherche. On désigne également ainsi le résultat de cette opération».
Le processus d'indexation vise à faciliter le repérage de l'information dans un corpus documentaire. En conséquence, les approches d'indexation utilisées doivent faire face à deux problèmes majeurs :

- Le choix des termes représentatifs de chaque document. En effet, le choix de la forme des descripteurs, des méthodes de pondération et de sélection de termes définit le schéma sous lequel le document sera présenté.
- L'évaluation des index et de leur pouvoir de représentation. La liste des index retenus devrait couvrir tout le document et bien décrire son contenu.

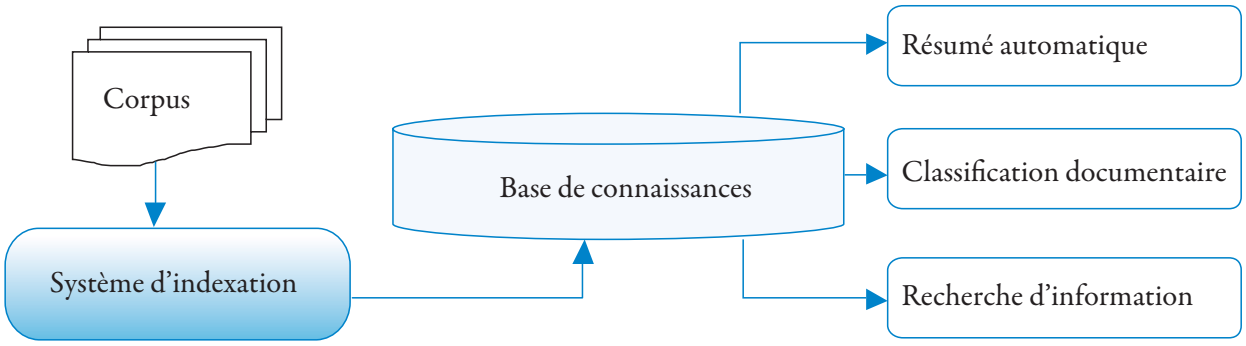

Figure 1. Processus de fouille de textes

2.1. Types d'indexation

L'indexation de toutes ses formes a pour but d'extraire les descripteurs les plus pertinents d'un document. Plus cette sélection est sophistiquée, plus les tâches ultérieures de fouille de textes exploitant le système d'indexation (classification, recherche d'information, etc.) s'avèrent précises. Il existe deux types fondamentaux d'indexation : l'indexation manuelle et l'indexation automatique.

En indexation manuelle, les descripteurs sont choisis par des experts ayant de bonnes connaissances linguistiques, le vocabulaire est alors contrôlé, une diversité de synonymes est utilisée et les aspects contextuels et sémantiques sont pris en considération. Malgré le fait que les spécialistes puissent se fonder sur les résumés des textes pour générer les index, l'indexation manuelle devient laborieuse dès que la masse documentaire s'accroît, et s'avère coûteuse en termes de temps de traitement des unités à indexer (Mallak, 2011).
L'indexation automatique, quant à elle, extrait des descripteurs automatiquement des textes en se fondant sur des règles d'analyse morphosyntaxique, des méthodes statistiques ou même sur des approches hybrides combinant les deux. Ce type d'indexation pallie les problèmes du coût de traitement, mais la conservation de la sémantique dans un document constitue alors un enjeu vital (Zaki, 2013).

2.2. Les formes de descripteurs

Les descripteurs sont les unités de texte qui représentent son contenu (Mallak, 2011). Il existe plusieurs catégories de descripteurs qui tentent d'une part de réduire la dimension du document, et d'autre part, de conserver son aspect sémantique. Nous en citons :

Les mots. Un mot est tout groupe de lettres formant un sens, compris entre deux séparateurs (espace, ponctuation, etc.). Le texte est alors segmenté en mots simples.

Les lemmes. Le processus de lemmatisation consiste à utiliser des règles grammaticales pour remplacer les verbes et les noms par leurs formes canoniques. Les lemmes peuvent ainsi correspondre à la forme des mots du dictionnaire.

Les racines *(stems)*. La racine est la plus petite unité lexicale qui permet de former un mot. Le processus de *stemming* extrait les racines de chaque mot du texte après l'élimination des préfixes, des infixes et des suffixes. Les mots partageant une racine commune sont tous associés à celle-ci.

Les concepts. Il s'agit des descripteurs issus d'une liste de lexique contrôlée, généralement des thésaurus ou des ontologies, et qui correspond aux notions principales contenues dans un document (Zaki, 2013).
Les multi-mots aussi appelés mots composés ou encore phrasèmes. Une séquence de mots décrit parfois plus précisément un document qu'un mot simple, en conservant mieux ses aspects contextuel et sémantique. Par exemple, l'expression «الأمم المتحدة» (Nations-Unies) est plus descriptive que «الأمم» (nations) et «المتحدة» (unies) prise séparément.

2.3. Préparation des documents

Avant d'entamer la phase d'indexation de documents arabes, une tâche très importante doit être accomplie, celle de nettoyage et de normalisation du texte (Bessou, Saadi *et al.*, 2007) (Mesleh, 2007a). Pour cette phase, les cinq étapes suivantes sont le plus souvent recommandées.

2.3.1. La segmentation (tokenization)

La segmentation est la production d'une séquence de segments séparés par des espaces ou des signes de ponctuation. La sortie est une liste de mots dépourvus de signe de ponctuation et de caractères spéciaux.

2.3.2. La suppression des mots vides

Les mots vides correspondent aux termes non porteurs d'information utile, qui figurent dans un texte. Il s'agit d'éliminer les mots dont l'occurrence est très fréquente et qui n'apportent aucune valeur ajoutée au processus d'indexation. Ces mots sont généralement des pronoms personnels, des articles ou des conjonctions, en l'occurrence : «أي», «أثناء», «إذ», etc. (Alajmi, Saad *et al.*, 2012).

2.3.3. Les conversions

La première conversion qui pourra être appliquée à un document arabe est l'élimination des signes diacritiques. Les signes diacritiques sont ajoutés au-dessus ou en dessous des lettres arabes afin de spécifier la prononciation du mot. Ce rôle phonologique influe aussi sur le sens de mot. En effet, deux mots peuvent être écrits de la même manière, mais différenciés par l'ajout de signes diacritiques différents. Par exemple, si le mot «عالم» est prononcé (عَالِم, *âalim*), il signifie «savant», et s'il est prononcé (عَالَم, *âalam*) il signifie «monde». Cette procédure vise à standardiser les documents du fait qu'il est rare de trouver un corpus entièrement accentué.
La deuxième conversion est celle des caractères qui a pour but de normaliser les lettres qui peuvent être écrites sous plusieurs formes. Ainsi les caractères «أ» «إ» et «آ» sont remplacés par «ا», de même «ة» est convertie en «ه» et «ي», «ئ» en «ى».

2.3.4. Le stemming

Le stemming consiste à extraire la racine d'un mot et à associer les mots liés morphologiquement à la même racine (Porter, 1980). Le nombre de termes est donc réduit, ce qui permet d'alléger le système. Cette technique dévoile un inconvénient majeur à savoir l'ambiguïté. Pour remédier à ce problème, la notion de *light stemming* est évoquée dans plusieurs travaux : elle consiste à éliminer juste les préfixes et les suffixes d'un mot donné, sans avoir à remonter à sa racine.

2.4. La représentation des documents

La représentation des documents est l'une des techniques qui sont utilisées pour réduire la complexité des documents et pour les rendre plus faciles à manipuler ; le document est alors transformé de sa version textuelle en une matrice [Document × Terme] (Figure 2). La représentation du document la plus utilisée est le modèle appelé vectoriel (VSM : *Vector Space Model*) dans lequel les documents sont représentés par des vecteurs de termes. Cette représentation a ses propres limites comme la grande dimension de représentation et la perte de corrélation entre les termes adjacents, ce qui entraîne la perte de la relation sémantique qui existe entre les termes d'un document. Pour surmonter ces problèmes, les méthodes de pondération sont utilisées pour attribuer des poids appropriés aux termes comme le représente la figure 2.

$$\begin{bmatrix} & \mathbf{T_1} & \mathbf{T_2} & \cdots & \mathbf{T_m} & \\ \mathbf{D_1} & p_{11} & p_{12} & \cdots & p_{1m} & \mathbf{C_a} \\ \mathbf{D_2} & p_{21} & p_{22} & \cdots & p_{2m} & \mathbf{C_b} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{D_n} & p_{n1} & p_{n2} & \cdots & p_{nm} & \mathbf{C_k} \end{bmatrix}$$

Figure 2. Matrice Document × Terme.

Chaque entrée représente un vecteur de termes où p_{nm} est le poids du terme $\mathbf{T_m}$ dans le document $\mathbf{D_n}$ et $\mathbf{C_i}$ est la classe attribuée au document $\mathbf{D_i}$.

2.5. La pondération

La pondération d'un terme d'indexation est l'association de valeurs numériques appelées poids à ce terme, de manière à représenter son pouvoir de discrimination pour chaque document de la collection. Cette caractérisation est liée au pouvoir informatif du terme pour le document donné. Le pouvoir de représentation d'un terme est parfois nommé l'informativité du terme. Cette notion fait référence à la quantité de sens qu'un mot porte.
Par exemple, la méthode TF-IDF (*Term Frequency – Inverse Document Frequency*) (Salton, Wong *et al.*, 1975) est l'une des méthodes les plus répandues dans le domaine de recherche documentaire (Trstenjak, Mikac *et al.*, 2013) (Zaki, Mammass *et al.*, 2010) (Hmeidi, Hawashin *et al.*, 2008) et elle est notamment très utilisée en modèle vectoriel.
TF représente le nombre d'occurrences d'un mot dans le document. IDF est la fréquence absolue inverse et égale à :

$$IDF_t = \log (N/n_t) \quad (1)$$

avec **N**, le nombre total de documents dans la collection et $\mathbf{n_t}$, le nombre de documents où le terme **t** apparaît.
Le poids d'un terme **t** dans le document d s'écrit généralement :

$$Poids_d(t) = TF_{dt} \times IDF_t \quad (2)$$

où $\mathbf{TF_{dt}}$ est la fréquence d'apparition du terme **t** dans le document **d** et $\mathbf{IDF_t}$ est la fréquence absolue inverse du terme **t** dans le corpus. Ainsi, le poids d'un terme augmente si celui-ci est fréquent dans le document et décroît si celui-ci est fréquent dans la collection.
Il existe d'autres façons de déterminer le poids d'un terme, en l'occurrence, la pondération booléenne, la fréquence de mots, l'entropie, etc. Cependant, les méthodes purement statistiques ont deux inconvénients majeurs. D'une part, il en résulte une énorme matrice creuse, ce qui pose un problème de grande dimension. D'autre part, elles ignorent la modélisation sémantique du document. De nouvelles méthodes que nous aborderons dans la suite sont apparues pour pallier ces limites.

2.6. La réduction de dimension

Après le prétraitement et l'indexation, une étape importante pour la classification de textes s'impose : il s'agit de réduire la dimension du texte (Mountassir, 2012). L'idée principale est de sélectionner un sous-ensemble de termes caractéristiques du document, et ce, en gardant les mots dotés des scores les plus élevés, en appliquant des mesures confirmant l'importance des termes sélectionnés. De nombreuses mesures d'évaluation des termes sont utilisées dans la littérature, nous en citons : le seuillage de fréquence (*Document Frequency Thresholding*), le gain d'information, la mesure de Chi-deux χ^2, *Odds Ratio* et l'information mutuelle.

2.7. La classification

La classification du texte est une partie importante du processus de fouille de textes (Figure 1). Elle consiste à fournir un ensemble de données d'apprentissage (documents étiquetés) au système de classification. La tâche est alors de déterminer un modèle de classification qui soit capable d'affecter la bonne classe à un nouveau document. Ces dernières années, la tâche de classification automatique de textes a été largement étudiée et les progrès semblent rapides dans ce domaine (Al-Mahmoud et Al-Razgan, 2015). Plusieurs méthodes de classification ont fait l'objet d'études comparatives et ont prouvé leur efficacité. A titre illustratif, nous citons : le classificateur bayésien, les arbres de décision, K-plus proche voisin (K-ppv), *Support Vector Machines* (SVM), et les réseaux de neurones (Alsaleem, 2011) (Bawaneh, Alkoffash *et al.*, 2008) (Alsalemi et Aziz, 2011) (El-Kourdi, Bensaid *et al.*, 2010).
La classification des documents textuels présente de nombreux défis et difficultés. Tout d'abord, il est difficile d'exprimer la sémantique de haut niveau et des concepts abstraits de la langue naturelle avec seulement quelques mots-clés, ce qui confirme le fait que l'efficience de l'étape d'indexation est primordiale et décisive.

2.8. Evaluation des systèmes d'indexation

L'évaluation expérimentale des classificateurs représente la dernière étape du processus d'indexation. Elle tente généralement d'évaluer l'efficacité d'un classificateur, à savoir sa capacité de prendre les décisions de catégorisation. Il existe à cet effet de nombreuses mesures, chacune mettant en évidence telle ou telle propriété du système. Nous avons retenu les mesures les plus utilisées suivantes : le rappel (3) qui est synonyme du taux de vraie acceptation, la précision (4) qui mesure le taux de bonnes réponses parmi les
réponses positives et la f-mesure (5) qui synthétise les deux premières. Considérons les nominations suivantes :

- **TP** *(True positive)* i.e. le nombre de documents correctement attribués à une catégorie,
- **FN** *(False Negative)* i.e. le nombre de documents incorrectement attribués à une catégorie,
- **FP** *(False positive)* i.e. le nombre de documents incorrectement rejetés affectés à une catégorie,
- **TN** *(True Negative)* i.e. le nombre de documents correctement rejetés attribués à une catégorie.

$$\textbf{Rappel} = \frac{TP}{TP + FN} \quad (3) \qquad \textbf{Précision} = \frac{TP}{TP+FP} \quad (4)$$

$$\textbf{F-mesure} = \frac{2 \times \text{Rappel} \times \text{Précision}}{(\text{Rappel} + \text{Précision})} \quad (5)$$

3. L'extraction des descripteurs

La fouille de textes repose sur un ensemble de techniques qui analysent de grandes quantités de données, extraient des relations qui sont inconnues au préalable, et fournissent des solutions pour aider à mieux représenter et exploiter ces données. L'indexation des documents, appelée aussi extraction des descripteurs, consiste à extraire les mots clés les plus pertinents dans un document qui décrivent mieux son contenu.

3.1. Problématique

Par rapport à d'autres langues, la langue arabe a une variation morphologique très riche et des caractéristiques syntaxiques extrêmement complexes, ce qui est l'une des principales raisons qui explique le manque de méthodes de recherche dans le domaine du traitement des textes arabes (El-Halees, 2007), (Samir, Ata *et al.*, 2005). L'indexation et la classification de textes sont des tâches importantes de ce traitement. Un processus typique de la classification de textes se compose des étapes suivantes : prétraitement, indexation, réduction de la dimension et classification (Wei, Gao *et al.*, 2010).

Un ensemble de modèles de classification et des techniques d'apprentissage automatique ont été appliqués à la classification de textes arabes, comme l'illustre la liste suivante :

- les K plus proches voisins (Kanaan, Al-Shalabi *et al.*, 2006) (Syiam, Fayed *et al.*, 2006),
- le modèle bayésien (El Kourdi, Bensaid *et al.*, 2004),
- SVM (Gharib, Habib *et al.*, 2009) (Alsaleem, 2011) (Mesleh, 2007b), (Mesleh, 2007c),
- les réseaux de neurones (Harrag, El-Qawasmah *et al.*, 2011),
- le maximum d'entropie (El-Halees, 2007) (Sawaf, Zaplo *et al.*, 2001),
- l'algorithme de Rocchio (Syiam, Fayed *et al.*, 2006),
- le classificateur à base de distances (Duwairi, 2005) (Khreisat, 2006) (Duwairi, 2006),
- les classificateurs à base de connaissances WordNet (Benkhalifa, Mouradi *et al.*, 2001).

Cependant, la phase d'extraction des descripteurs n'a pas eu le même intérêt, malgré son rôle primordial et décisif en classification. (Al-Mahmoud et Al-Razgan, 2015) présentent une étude systématique des techniques de fouille de textes qui confirme le manque de travaux concernant l'extraction des caractéristiques des documents arabes. La figure 3 illustre la distribution des techniques de fouille de textes décrites dans cette étude qui a couvert plus de cent articles.

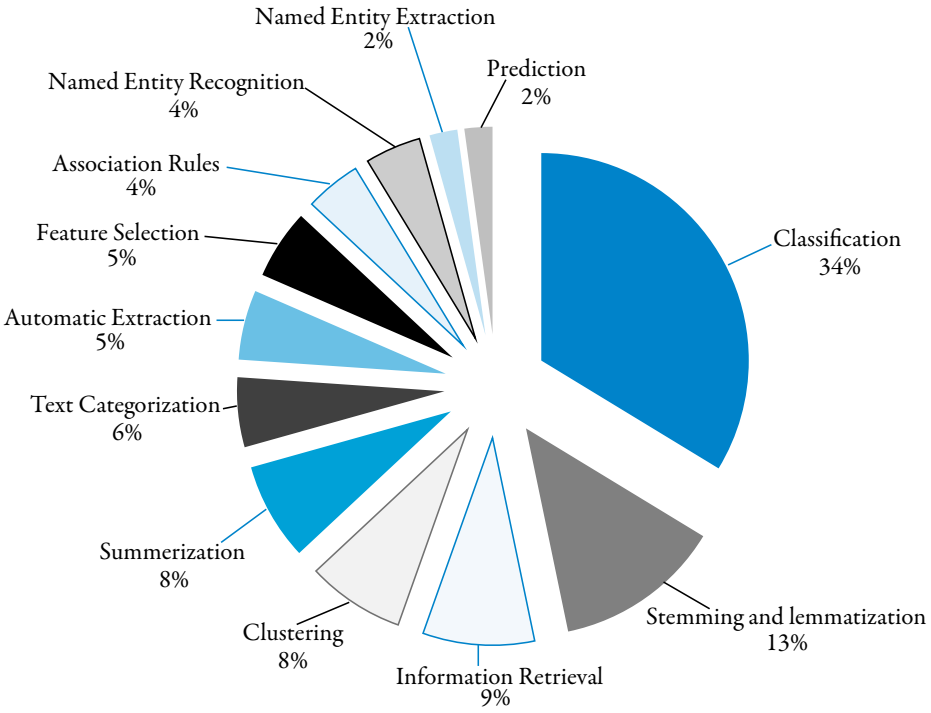

Figure 3 : Distribution des études en fouille de textes arabes selon (Al-Mahmoud et Al-Razgan, 2015).

Dans cet article, nous avons adopté une sélection qualitative des articles. Ainsi, avons-nous retenu les travaux montrant des contributions remarquables au niveau de l'extraction des descripteurs et présentant des résultats considérables.

3.2. Méthodes d'extraction des descripteurs

Dans cette section, nous souhaitons, d'une part, catégoriser les travaux selon les approches les plus utilisées pour synthétiser les avancées concernant l'extraction des descripteurs. D'autre part, nous introduisons quelques travaux susceptibles d'inspirer un système d'indexation de la langue arabe plus sophistiqué.

3.2.1. Approches linguistiques

Une approche linguistique consiste à apporter une analyse morphologique et syntaxique profonde du document traité, et ce, en se fondant sur les règles grammaticales et les relations entre les différentes unités textuelles, pour des fins de désambiguïsation sémantique, ou plus encore, d'indexation. Plusieurs tentatives de modélisation des règles linguistiques ont été proposées dans la littérature.

Al Molijy *et al.* (2012) utilisent l'analyse syntaxique des mots du document. L'algorithme proposé consiste à découper les mots en N-grammes (où N vaut 3, 4 ou 5), calculer leurs fréquences, ensuite retenir les 100 premiers mots les plus fréquents constituant ainsi un profil N-gramme du document. Cette méthode, utilisée aussi en processus de *stemming*, permet de réduire le nombre des mots représentant un document.

Mansour *et al.* (2008) procèdent à une analyse morphologique des mots du document pour extraire les index. D'une part, les auteurs proposent un processus d'extraction des stems. D'autre part, ils mettent en place un système de reconnaissance des noms et des verbes en se fondant sur les rimes et les règles grammaticales. Un poids est ensuite attribué à chaque stem en tenant compte de son occurrence et en introduisant une fonction indiquant comment le mot est étalé dans le document.
Aussi bien (Saadane, 2013) que (Bessou, Saadi *et al.*, 2007) proposent des systèmes d'extraction des connaissances, fondés sur une analyse linguistique profonde et faisant appel à une ontologie de domaine pour révéler le contenu sémantique. Les résultats de leurs travaux s'annoncent prometteurs, mais révèlent d'autres problématiques nécessitant des études minutieuses.
Quant à Hulth (2003), il présente une approche d'extraction de *chunk* nominaux (unité textuelle minimale ayant un sens, composée d'un mot ou plus) et de N-gramme. Dans ses expériences, Hulth intègre l'étiquetage en parties de discours, ce qui augmente significativement les performances du système.
Ces méthodes sont largement exploitées en fouille de textes arabes, grâce à la précision de leurs résultats et la fiabilité des algorithmes de reconnaissance syntaxique et sémantique. Le défi majeur de ces méthodes est de couvrir la diversité grammaticale et le vocabulaire de la langue arabe.

3.2.2. Approches numériques

Ces approches se fondent sur les techniques statistiques, la théorie des graphes ou les approches sémantiques prises séparément ainsi que sur leur combinaison.

3.2.2.1. Méthodes statistiques

Cette section du papier présente les travaux qui sont fondés sur des méthodes et des mesures purement statistiques pour l'extraction des mots clés. Ceci est le critère que nous avons adopté pour préparer ce regroupement de méthodes.
Nombreux sont les travaux qui ont adopté des approches statistiques pour l'extraction des mots clés, en étudiant le comportement des termes candidats dans un document, voire dans le corpus. Plus un terme candidat est jugé important dans le document, plus celui-ci est pertinent comme terme clé.
La méthode TF-IDF (5) fournit une bonne représentation du poids pour les corpus dont les documents sont de tailles homogènes, c'est-à-dire composés de documents de tailles similaires.
De nombreuses variantes de TF-IDF sont proposées dans la littérature, et elles ont fait objet d'un grand nombre de comparaisons. Okapi qui est une méthode alternative à la TF-IDF est très utilisée en recherche d'information. Elle prend mieux en compte la longueur des documents (Robertson, Walker *et al.*, 2000) (6) (7).

$$\text{Okapi (terme)} = \text{TF}_{\text{BM25}}\,\text{(terme)} \times log\left(\frac{N\text{-}DF(terme)+0{,}5}{DF(terme)+0{,}5}\right) \quad (6)$$

$$\text{avec :}\ \ \text{TF}_{\text{BM25}} = \frac{TF(terme)\times(k1+1)}{TF(terme)+k1\times\left(1-b+b\times\frac{DL}{DLmoyenne}\right)} \quad (7)$$

où **DL** représente la longueur du document traité et $\mathbf{DL}_{\mathbf{moyenne}}$ la longueur moyenne des documents de la collection. $\mathbf{K}_1$ et **b** sont des constantes fixées respectivement à 2 et 0,75 (Bougouin, 2013).
El-Khoribi et Ismael (2006) ont appliqué les stems comme caractéristiques de représentation. Ces caractéristiques sont ensuite représentées en tant que vecteurs de dimension égale au nombre de classes où la probabilité d'appartenance d'un stem est prise en considération. Ensuite une table de correspondance de stems est construite à partir des racines et des étiquettes des classes auxquelles elles appartiennent. Après, le modèle de Markov caché (HMM : Hidden Markov Model) est utilisé pour évaluer l'appartenance d'un nouveau document à une classe.
Khreisat (2006) a construit un système de classification de documents textuels arabes à l'aide de la technique statistique fréquentielle N-grammes et en utilisant une mesure de dissemblance appelée la distance de Manhattan, et l'opérateur de Dice comme mesure de similarité. La mesure de Dice a été utilisée à des fins de comparaison. Les résultats ont montré que la classification de textes en utilisant les N-grammes et la mesure de Dice surpasse la classification fondée sur les N-grammes et la mesure de Manhattan.
El-Halees (2007) a présenté des résultats prometteurs obtenus en utilisant des méthodes statistiques telles que l'entropie maximale sur une base d'articles en arabe et sans analyse morphologique.
(Mesleh, 2007a) a étudié l'usage du classificateur SVM avec six techniques de sélection des caractéristiques. Leurs expériences montrent que χ^2 s'impose par rapport aux autres
techniques.
Dans (Thabtah, Hadi *et al.*, 2008), les variantes du modèle vectoriel sont étudiées à l'aide de l'algorithme K-ppv. Ces variantes sont le coefficient Cosinus, le coefficient de Dice et le coefficient de Jacaard, en utilisant différentes méthodes de pondération des termes. Les résultats obtenus sur une base arabe ont montré que les performances

obtenues par Dice-TFIDF et Jaccard-TFIDF surpassent celles obtenues par Cosinus TFIDF, Cosinus à base FDIF, Cosinus-ITF, Cosinus à base log(1 + *TF*), Dice à base FDIF, Dice à base ITF, Dice à base log(1+*TF*), Jaccard à base FDIF, Jaccard à base ITF, et Jaccard à base log(1+*TF*).
C'est ACO *(Ant Colony Optimization)* qui est appliqué dans (Mesleh et Kanaan, 2008) comme mécanisme de réduction de l'espace des caractéristiques. La méthode χ^2 est utilisée comme une fonction de calcul de scores. Ils ont ensuite procédé à la classification des documents arabes en utilisant le classificateur SVM.
(Al-Shalabi et Obeidat, 2008) ont utilisé un K-ppv pour classer les documents arabes. Ils extraient en tant que caractéristiques des mots clés donnés par les unigrammes et les bigrammes, ensuite la mesure de TFIDF est appliquée en tant que procédé de sélection de ces caractéristiques.
(Al-Harbi, Almuhareb *et al.*, 2008) a testé le SVM et la C5.0 sur sept corpus arabes avec des descripteurs pondérés par χ^2. Les performances obtenues sont de 86% pour le SVM et de 92% pour C5.0.
(Bawaneh, Alkoffash *et al.*, 2008) a comparé les deux classificateurs K-ppv et NB. Le *light stemmer* a été utilisé comme caractéristique et la mesure TFIDF en tant que méthode de pondération des caractéristiques. Le classificateur K-ppv a été jugé plus performant.
(Thabtah, Eljinini *et al.*, 2009) a mis en place un système de catégorisation arabe en utilisant le classificateur bayésien naïf fondé sur les caractéristiques de pondération fournies par le test de χ^2 pour classer une simple base de données étiquetées. Les résultats expérimentaux montrent que la sélection des caractéristiques améliore souvent la précision de la classification en supprimant les termes vides ou rares.
Dans (Kanaan *et al.*, 2009), les documents en arabe sont classés avec l'algorithme espérance-maximisation (EM). La mesure TFIDF est appliquée en tant que méthode de pondération des éléments caractéristiques tandis que l'algorithme bayésien naïf est utilisé pour calculer les étiquettes des documents et que finalement on procède à la classification en utilisant l'algorithme EM.
(Zubi, 2009) a comparé les deux classificateurs K-ppv et NB appliqués à une base de 1562 documents. Ces derniers sont classés en 6 catégories et pondérés en utilisant la mesure TFIDF. L'expérience a montré que K-ppv est plus performant.
(Gharib *et al.*, 2009) ont appliqué quatre classificateurs, SVM, Bayésien naïf, K-ppv et la méthode de Rocchio, à une base de documents arabes, en utilisant le *stemming* comme méthode de représentation des caractéristiques et la mesure TFIDF comme méthode de pondération. Le classificateur de Rocchio fonctionne mieux lorsque l'espace des caractéristiques est petit, mais le SVM est plus performant quand l'espace devient de plus en plus grand.
Dans (Al-Shalabi, Kanaan *et al.*, 2010), l'algorithme de k plus proches voisins et les mots clés sont extraits selon leur pondération TFIDF dans les documents, en obtenant une micro-moyenne précision de 95%.
Dans leur étude comparative (Raheel et Dichy, 2010) ont montré l'influence du choix de type d'entités à manipuler sur les performances des classificateurs. Ils ont choisi comme descripteurs, les mots dans leur forme originale, les lemmes, les racines, et les n-grammes. Deux classificateurs ont été utilisés, le SVM et les réseaux bayésiens naïfs. Le SVM basé sur les 3-grammes a donné de meilleurs résultats de classification avec une F-mesure dépassant 92%.
(Al-Salemi et Aziz,2011) ont utilisé des techniques de sélection de caractéristiques telles que l'information mutuelle, la mesure statistique χ^2, le gain d'information, le coefficient ESG et Odds Ratio pour réduire la dimension de l'espace des caractéristiques en éliminant les éléments qui sont considérés comme non pertinents pour une catégorie étudiée.
D'autres modèles sont utilisés dans la littérature (Hasan et Ng, 2010) (Bougouin, 2013) (Mesleh, 2007a) tels que LSI *(Latent Semantic Indexing)* qui prend en considération la sémantique des termes pour la représentation des documents. Les documents sont représentés dans un espace réduit de termes d'indexation. (Hofmann, 1999) propose un modèle probabiliste de *Latent Semantic Indexing* (PLSI). Il émet l'hypothèse que les documents sont associés à un certain nombre de sens et que les termes correspondent à l'expression de ces sens.
En conclusion, ces méthodes, considérées comme simples à implémenter, sont efficientes et parfaitement tolérantes aux grandes masses documentaires. D'autre part, l'hypothèse considérant les mots comme étant des unités indépendantes engendre une perte d'information sémantique. Les index qui en résultent peuvent générer des problèmes de polysémie et dévier du contexte général du document.

3.2.2.2. Méthodes fondées sur les graphes

Ces méthodes proposent de représenter le texte sous forme de graphe. Généralement, les mots constituent les nœuds du graphe et les arcs représentent la relation entre les mots (relation sémantique, structurelle, etc.).
Mihalcea et Tarau (2004) proposent l'algorithme TextRank, une adaptation textuelle de l'algorithme PageRank (Page, Brin *et al.*, 1998). Il consiste à représenter les documents textuels sous forme de graphe où les nœuds peuvent représenter un mot ou un groupe de mots. Une pondération $\mathbf{w_{n1n2}}$ est associée à chaque arc liant deux nœuds $\mathbf{n_1}$ et $\mathbf{n_2}$, et représente la fréquence de cooccurrence des deux termes dans une fenêtre de **N** mots.
Le score du nœud $\mathbf{n_i}$, noté $\mathbf{S(n_i)}$, est initialisé par une valeur par défaut, et il est ensuite calculé d'une manière itérative jusqu'à convergence en utilisant la formule suivante :

$$S(n_i)=(1-d)+ d \times \sum_{n_j \in Adj(n_i)} \frac{w_{ji}}{\sum_{n_k \in Adj(n_j)} w_{jk}} S(n_i) \quad (8)$$

où **Adj(n_j)** représente les voisins de $\mathbf{n_i}$, et **d** est un facteur d'amortissement fixé à 0.85 (Page, Brin *et al.*, 1998). Intuitivement, un nœud recevra un score élevé si ses voisins ont des scores élevés. Finalement, après convergence, les **k%** termes ayant des scores élevés sont élus comme mots clés.

Dans leurs travaux, Mihalcea et Tarau utilisent l'étiquetage en parties de discours afin de réduire la liste des termes représentés par le graphe, et ce, en ne considérant que les noms et les adjectifs, ce qui améliore les performances du système. Cependant, l'orientation du graphe n'apporte pas d'amélioration à considérer, par rapport au graphe non orienté.

Des variantes de cet algorithme ont vu le jour, par exemple (Wan et Xiao, 2008b). Les auteurs proposent l'algorithme SingleRank qui définit trois différences majeures par rapport à TextRank. Premièrement, les arcs ont des poids correspondants au nombre de cooccurrence des deux termes connexes alors que les arcs ont le même poids pour TextRank. D'autre part, TextRank procède à un filtrage de termes contrairement à SingleRank qui n'effectue aucune discrimination. En outre, pour chaque phrasème candidat, le score est calculé en sommant les scores des termes formant ce phrasème, obtenus de la représentation graphe SingleRank. Les phrasèmes candidats ayant les plus grands scores sont considérés des termes clés.

ExpandRank (Wan et Xiao, 2008b) est une extension de TextRank qui consiste à exploiter le voisinage du document analysé. Pour un document **d**, les k-plus proches voisins sont trouvés à partir des documents de la collection, le graphe est généré ensuite à partir du document traité et ses k plus proches voisins. Ainsi, chaque document $\mathbf{d_0}$ est-il réuni avec ses documents voisins dk, formant un document plus large $\mathbf{d_{k+1}}$ qui servira à la construction du graphe. Les termes candidats correspondent aux nœuds et un arc relie deux nœuds si les termes candidats co-occurrent dans une fenêtre de N mots du document. Le poids de l'arc liant deux nœuds $\mathbf{v_i}$ et $\mathbf{v_j}$ est donné par :

$$w(v_i, v_j) = \sum_{dk \in D} sim(d_0, dk) \times freq_{dk}(v_i, v_j) \quad (9)$$

où :

- **sim($\mathbf{d_0}$, $\mathbf{d_k}$)** est la similarité cosinus entre $\mathbf{d_0}$, et $\mathbf{d_k}$,
- **freqdk**($\mathbf{v_i}$, $\mathbf{v_j}$) est la fréquence de cooccurrence des termes $\mathbf{v_i}$ et $\mathbf{v_j}$ dans dk.

Dès que le graphe est construit, le reste de la procédure est similaire à SingleRank.

MedRank est un algorithme proposé par (Herskovic et Jorge, 2011) pour réordonner les rangs des concepts extraits d'une base médicale. Ces concepts sont extraits par le programme MetaMap dans un premier temps. De nouveaux scores sont ensuite affectés aux concepts en utilisant l'algorithme TextRank. Les meilleurs résultats sont obtenus en utilisant l'approche MedRank.

	Référence	Techniques utilisées	Observation
Méthodes Statistiques	(El-Khoribi et Ismael, 2006)	probabilité d'appartenance, HMM	HMM est utilisé pour évaluer l'appartenance d'un nouveau document à une classe
	(Khreisat, 2006)	N-grammes, la distance de Manhattan, la mesure de Dice	La classification de textes en utilisant les N-grammes avec la mesure de Dice surpasse la classification en utilisant les N-grammes avec la mesure de Manhattan
	(El-Halees, 2007)	l'entropie maximale	Les résultats obtenus sont prometteurs.
	(Mesleh, 2007a)	χ^2, SVM.	Six techniques de sélection des caractéristiques introduites. χ^2 est la plus performante.
	(Thabtah, Hadi *et al.*, 2008)	K-ppv, Cosinus, Dice, Jaccard, TFIDF.	Les résultats obtenus sur une base arabe ont montré que les performances obtenues par Dice-TFIDF et Jaccard-TFIDF sont les plus élevées.
	(Mesleh et Kanaan, 2008)	Ant Colony Optimization (ACO), χ^2, SVM.	ACO est appliqué pour réduire l'espace de représentation des caractéristiques et la méthode χ^2 pour le calcul de scores.
	(Al-Shalabi et Obeidat 2008)	K-ppv, unigrammes, bigrammes, TFIDF	Les mots clés donnés par les unigrammes et les bigrammes, sont pondérés par la mesure de TFIDF.
	(Al-Harbi, Almuhareb *et al.*, 2008)	SVM, C5.0, χ^2.	Le test est effectué sur sept corpus arabes.

Méthodes Statistiques *(suite)*	(Bawaneh, Alkoffash *et al.*, 2008)	K-ppv, NB, TFIDF	Le light stemmer a été utilisé comme caractéristique.
	(Thabtah, Eljinini *et al.*, 2009)	NB, χ^2.	Les résultats expérimentaux montrent que la sélection des caractéristiques améliore souvent la précision de la classification.
	(Kanaan *et al.*, 2009)	espérance-maximisation, TFIDF, NB.	Les documents arabes sont classés avec l'algorithme espérance-maximisation (EM).
	(Zubi, 2009)	K-ppv, NB, TFIDF.	La classification se fait sur un corpus de 1562 documents appartenant à 6 catégories différentes. L'expérience a montré que K-ppv est plus performant.
	(Gharib *et al.*, 2009)	SVM, NB, K-ppv, Rocchio, TFIDF	Le classificateur de Rocchio fonctionne mieux lorsque l'espace des caractéristiques est petit mais le SVM est plus performant quand l'espace devient grand.
	(Al-Shalabi, Kanaan *et al.*, 2010)	K-ppv, TFIDF.	Implémentation classique pour la classification des textes arabes.
	(Raheel et Dichy, 2010)	SVM, NB, 3-grammes.	Dans leur étude comparative, les auteurs ont montré l'influence du choix de type d'entités à manipuler sur les performances des classificateurs.
	(Al-Salemi et Aziz,2011)	l'information mutuelle, χ^2, le gain d'information, le coefficient ESG et Odds Ratio.	Ces techniques sont utilisées pour réduire la dimension de l'espace des caractéristiques en éliminant les éléments qui sont considérés comme non pertinents pour une catégorie étudiée.
Méthodes fondées sur les graphes	(Mihalcea et Tarau, 2004)	TextRank.	Adaptation textuelle de l'algorithme PageRank. Les nœuds peuvent représenter un mot ou un groupe de mots et les arcs n'importe quelle relation reliant les mots.
	(Wan et Xiao, 2008b).	SingleRank.	Variantes de TextRank.
	(Wan et Xiao, 2008b)	ExpandRank, K-ppv.	Extension de TextRank qui consiste à exploiter le voisinage du document analysé.
	(Herskovic et Jorge, 2011)	MedRank, MetaMap,	MedRank est un algorithme pour réordonner les rangs des concepts extraits d'une base médicale en utilisant l'algorithme TextRank

Tableau 1 : Synthèse des approches numériques.

Les méthodes d'indexation à base de graphes semblent être mieux adaptées aux textes bruts pour leur efficacité à conserver l'aspect structurel. Cependant, la complexité de calcul des scores des nœuds générés à partir des textes constitue une limite majeure. Pourtant, les méthodes statistiques restent les plus utilisées pour leur simplicité en implémentation et leurs résultats efficients.

3.2.2.3. Approches sémantiques

Ces approches visent, d'une part, à lever l'ambiguïté sur le sens des mots et d'autre part, elles permettent de tisser les relations sémantiques entre ces mots. Les textes sont représentés par des concepts symbolisant le sens plutôt que des mots simples. Les relations sémantiques peuvent aussi être calculées par le biais des méthodes évaluant la quantité d'information entre les mots deux à deux, en l'occurrence, l'information mutuelle.

D'autres chercheurs sont allés jusqu'à l'exploration de l'information contextuelle. Le travail de Zargayouna et Salotti (2004), testé sur un corpus de documents semi-structurés en XML, considère le document comme un ensemble d'unités sémantiques (les balises) représentant chacune un contexte particulier d'occurrence des termes. Néanmoins, dans (Roche, 2011), l'auteur travaille sur la désambiguïsation des acronymes et définit le contexte comme des mots caractéristiques présents dans la page dans laquelle l'acronyme à définir est présent. Pareillement, Motasem et Joseph (2009) proposent une méthode d'exploration contextuelle afin de lever l'ambigüité de la séquence «alif-noun».
Cependant, Jamoussi (2009) propose une méthode d'extraction de mots clés en se fondant sur la représentation sémantique des termes. Il présente deux mesures fondées sur des distances sémantiques, la distance de Kullback-Leibler (DKL) et l'information mutuelle moyenne (IMM), pour calculer la quantité d'information entre deux mots ou deux classes de mots. Cette méthode est testée par rapport à une représentation vectorielle simple, avec trois classificateurs non supervisés : l'algorithme K-means, les cartes de Kohonen et le réseau bayésien AutoClass. Le Tableau 2 synthétise les résultats obtenus. Ces résultats expriment, en pourcentage, le taux de bonne classification, en précisant les intervalles de confiance. La performance des résultats met en évidence l'importance de l'utilisation des mesures sémantiques pour la fouille de textes.

Méthode Jamoussi		K-means	Kohonen	AutoClass
Représentation vectorielle simple	*DKL*	70,5 ± 2,2	75,5 ± 2,1	81,3 ± 1,9
	IMM	74,1 ± 2,1	77,1 ± 2,1	84,7 ± 1,8
Représentation matricielle mixte	*DKL*	72,5 ± 2,2	76,7 ± 2,1	86,3 ± 1,7
	IMM	76,4 ± 2,1	80,4 ± 1,9	89,3 ± 1,5

Tableau 2 : L'apport de l'approche matricielle mixte par rapport à la représentation vectorielle standard (Jamoussi, 2009)

En outre, une autre technique originale est exploitable pour les données textuelles en langue arabe, il s'agit du regroupement sémantique. (Liu, Peng *et al.*, 2009) proposent une méthode d'extraction de mots clés fondée sur le regroupement sémantique qui garantit une bonne couverture sémantique du document. La méthode extrait les termes candidats qui seront regroupés en classes après le calcul des liens sémantiques entre ces termes. Ce regroupement consiste à élaborer un ensemble de mots de référence pour chaque classe. Les mots de référence sont utilisés pour l'extraction des mots clés après le filtrage des termes candidats. Un mot clé doit contenir au moins un mot de référence.

3.2.3. Approches hybrides

L'adoption des approches hybrides pour la fouille de données textuelles est devenue courante. Plusieurs chercheurs essaient différentes combinaisons des méthodes linguistiques, numériques et sémantiques afin de révéler l'information cachée dans un document, et enrichir les liens contextuels qu'il contient (Bessou, Saadi *et al.*, 2007) (Jamoussi, 2009) (Saadane, 2013). Ces approches aboutissent souvent à des résultats meilleurs que ceux obtenus par l'utilisation des méthodes standards.
Par exemple, Tomokyo et Hurst (2003) proposent une méthode qui vérifie la grammaticalité (un mot clé doit être bien formé syntaxiquement) et l'informativité (le mot clé doit exprimer au moins une idée du contexte général du document) en utilisant la Kullback-Leibler divergence. Ainsi, pour un terme candidat, plus sa probabilité de passer du modèle uni-gramme généré à partir du document analysé au modèle N-gramme généré par le même document augmente, plus il respecte la propriété de grammaticalité. De même, plus sa probabilité de passer du modèle N-gramme généré à partir d'un corpus de référence vers un modèle N-gramme généré par le document traité augmente, plus le terme candidat est informatif.
Dans le but de réduire l'espace des caractéristiques (Harrag, El-Qawasmah *et al.*, 2011) compare trois techniques de prétraitement : *light stemming*, *root-based stemming* et *dictionary lookup stemming*. Ensuite, deux classifieurs ont été testés : les réseaux de neurones artificiels (ANN) et le SVM. Les performances données par SVM sont supérieures à celles de ANN avec le *light stemming*.
La méthode de (Motasem et Joseph, 2009) se fonde essentiellement sur l'analyse morphosyntaxique et l'exploitation des règles grammaticales pour la reconnaissance des mots adjacents à la séquence en question, pour découvrir son contexte. Néanmoins, d'après (Alwedyan *et al.*, 2011), leur propre classificateur multi-classes à base de règles d'associations fonctionne mieux que NB et SVM.
Quant au travail (Zaki, Mammass et al, 2014), les auteurs introduisent la notion de voisinage sémantique. Ils proposent un système hybride pour l'indexation contextuelle et sémantique des documents arabes, apportant une amélioration aux modèles classiques fondés sur les n-grammes et le modèle Okapi. Ils calculent la similarité entre les mots en utilisant une hybridation de mesures statistiques N-grammes, Okapi et une fonction noyau. Afin d'avoir

un indice de descripteur robuste, ils ont utilisé un graphe sémantique pour modéliser les connexions sémantiques entre les termes, en s'appuyant sur un dictionnaire auxiliaire pour augmenter la connectivité du graphe. Tout d'abord, le document est modélisé par un graphe. Ensuite, le graphe est renforcé par un dictionnaire de concepts. Les pondérations des mots sont ensuite calculées en utilisant une fonction à base radiale (ABR). Ceci a permis d'améliorer les performances du système d'indexation. Le k-ppv est utilisé pour la classification. Le rappel et la précision sont adoptés comme métriques d'évaluation. Le tableau 3 illustre les résultats de cette approche en la combinant avec des méthodes d'indexations très utilisées.

Méthode	Corpus	Précision	Rappel
TF IDF	*Sport*	0.83	0.73
	Politique	0.68	0.61
	Economie	0.56	0.71
TF IDF + ABR	*Sport*	0.94	0.77
	Politique	0.78	0.67
	Economie	0.59	0.71

Méthode	Corpus	Précision	Rappel
Okapi	*Sport*	0.82	0.75
	Politique	0.79	0.67
	Economie	0.73	0.65
Okapi + ABR	*Sport*	0.91	0.81
	Politique	0.81	0.70
	Economie	0.76	0.69

Tableau 3. Impact de la méthode ABR combinée avec TFIDF et Okapi (Zaki, Mammass et al, 2014)

D'autres hybridations sont introduites afin d'améliorer les résultats de classification des documents arabes. (Raheel, Dichy *et al.*, 2009) a combiné la méthode de Boosting et l'arbre de décision comme classificateur hybride. Ils ont utilisé les lemmes comme formes de caractéristiques et la TFIDF pour la pondération. Une comparaison de la méthode a été faite avec deux classificateurs, Bayésien naïf (NB) et SVM. Les résultats montrent que SVM et NB surpassent l'approche proposée.

Le modèle SVM a montré des succès remarquables dans la classification de textes (Joachims, 1998). À cet effet, dans (Shafiei, Wang *et al.*, 2007), une méthode hybride utilise le TSVM *(Transductive Support Vector Machines)* et le recuit simulé (SA : *simulated annealing*). Les auteurs ont choisi les deux mille meilleures caractéristiques à partir du test de χ^2 pour former les données de base et ils ont obtenu de meilleurs résultats de classification avec SA par rapport aux SVM et TSVM.

Dans (Mohamed et Watada, 2010), l'analyse sémantique latente (LSA) produit une évaluation de chaque terme dans un document, puis à l'aide du raisonnement probant (ER : *Evidential reasoning*) une catégorie selon la base documentaire est assignée au nouveau document. Des expériences ont été effectuées sur une combinaison de ER avec la LSA et de ER avec TFIDF qui ont montré que ER-LSA est plus performant que ER-TFIDF.

(Zaki, Mammass *et al.*, 2010) étendent le modèle vectoriel en combinant la TF-IDF avec la formule Okapi pour l'extraction des concepts pertinents qui représentent un document. Il propose une nouvelle mesure qui prend en considération la notion de voisinage sémantique en utilisant une mesure de similarité entre termes, et en combinant le calcul du TF-IDF-okapis avec une approche noyau (fonction à base radiale). Cette approche d'indexation permet de valoriser la notion de proximité sémantique. Les résultats expérimentaux confirment l'apport de la contribution.

Chantar *et al.* (Chantar et Corne, 2012) ont proposé une méthode de sélection de caractéristiques appelée Binary Particle Swarm Optimization and K-ppv (OSPB K-ppv). Trois algorithmes d'apprentissage sont utilisés: le SVM, Bayes Naïf et l'arbre de décision C4.5, afin de classer les documents en langue arabe. Les résultats obtenus par SVM ainsi que Naïve Bayes montrent que OSPB K-ppv fonctionne bien en tant que technique de sélection de caractéristiques.

Référence	Linguistique	Statistique	Graphe	Sémantique	Commentaire
(Tomokyo et Hurst 2003)	√	√	x	x	Introduction des notions de grammaticalité et informativité.
(Harrag, El-Qawasmah *et al.*, 2011)	√	√	x	√	Trois techniques de prétraitement sont comparées: light stemming, root-based stemming et dictionary lookup stemming.
(Motasem et Joseph, 2009)	√	x	x	√	Utilisation de l'analyse morphosyntaxique et l'exploitation des règles grammaticales pour la reconnaissance des mots.
(Alwedyan *et al.*, 2011)	√	√	x	x	Introduction de classificateur multi-classes à base de règles d'associations

(Zaki, Mammass et al, 2014)	x	√	√	√	utilisation des méthodes statistiques et des graphes pour l'indexation contextuelle et sémantique des documents arabes
(Raheel, Dichy *et al.*, 2009)	√	√	x	x	Combinaison de la méthode Boosting et l'arbre de décision comme classificateur hybride et TFIDF pour la pondération.
(Shafiei, Wang *et al.*, 2007)	x	√	x	x	La méthode hybride utilisée est le TSVM (Transductive Support Vector Machines) avec (SA : simulated annealing).
(Mohamed et Watada, 2010)	x	√	x	√	Utilisation de l'analyse sémantique latente (LSA) avec (ER : Evidential reasoning). Des expériences ont été effectuées sur une combinaison de ER avec la LSA et de ER avec TFIDF qui ont montré que ER-LSA est plus performant que ER-TFIDF.
(Zaki, Mammass *et al.*, 2010)	x	√	x	√	Combinant du calcul de TFIDF-Okapis avec une fonction à base radiale.
(Chantar et Corne, 2012)	x	√	x	x	Chantar *et al.* ont proposé une méthode de sélection de caractéristiques appelée Binary Particle Swarm Optimization and K-ppv

Tableau 4 : Synthèse des méthodes hybrides.

L'extraction automatique des descripteurs à partir de données textuelles afin de les exploiter implique la mise en place en place d'un moyen de réduction de calcul et d'accélération de traitement, en particulier pour de grandes quantités de données, tout en étant efficace. Ainsi, différentes approches et méthodologies pour la modélisation et la représentation de données textuelles ont été proposées. Nous discuterons dans la section suivante les avantages et les inconvénients de chaque approche.

4. Discussion

Les approches statistiques représentent le texte en sac-de-mots. Cette représentation est adaptée pour capturer les fréquences d'apparition d'un mot et ignore l'information structurelle et sémantique que contient le document.
La représentation fondée sur les graphes est plus adéquate pour modéliser la structure et la sémantique, et son utilisation pour l'extraction de mots clés a prouvé son efficacité (Mihalcea et tarau, 2004). Or, une limite principale de la représentation fondée sur les graphes est la complexité du graphe généré pour chaque document, et le calcul des scores qui augmente d'une façon exponentielle relativement à la taille du document.
Les approches linguistiques, quant à elles, exploitent des règles morphosyntaxiques pour extraire les termes. Ce genre de techniques offre de bons résultats dans des cas spécifiques, en désambigüisation de mots par exemple, mais s'avère moins compétitif pour les systèmes d'indexation, vu la complexité de la langue en question.
Les méthodes utilisant des ressources sémantiques externes (dictionnaire, ontologie ou autres) offrent une meilleure couverture sémantique du document. Sauf que la reconnaissance des unités sémantiques reste limitée au domaine décrit par la ressource utilisée. La génération automatique des dictionnaires à partir du corpus étudié s'avère être une piste prometteuse.
D'autre part, plusieurs chercheurs essaient différentes combinaisons des méthodes classiques, introduisant ainsi des méthodes hybrides de traitement de documents. Ces méthodes permettent non seulement d'augmenter le résultat d'indexation, mais elles deviennent aussi indispensables dans des cas de traitement particulier. Le tableau 5 synthétise les avantages et les limites de chacune de ces approches.
La fouille des documents en langue arabe est confrontée à un autre problème, celui des évaluations des méthodes sur les corpus. Dans la plupart des travaux sur les textes arabes, et en l'absence d'un corpus standard libre, les auteurs construisent leurs propres corpus. Ils choisissent le nombre des catégories et leurs thèmes. Pour chaque catégorie, les documents sont collectés manuellement. Les documents appartenant à plusieurs catégories sont souvent éliminés. Or, pour tester la précision des différentes méthodes, elles doivent être appliquées au même corpus. Plus encore, pour qu'une méthode prouve son efficacité, elle doit être appliquée à plusieurs corpus de taille et de thèmes différents. Un corpus standardisé encouragera donc les auteurs à introduire des nouvelles méthodes, comparer l'efficacité des approches d'une façon objective, et avoir une synthèse plus significative de l'état de l'art.

Approches		Avantages	Inconvénients
Approche linguistique		• Efficacité importante au niveau sémantique • Bon rapport pertinence / représentation	• Laborieuse en cas de grandes masses documentaires • Difficile de prendre en charge toute la complexité de la langue arabe • Relative à la langue
Approche Numérique	***Méthodes statistiques***	• Simples à déployer • Grande progression lors des dernières années • Résultats considérables du point de vue mathématique	• Les mots sont considérés indépendants (sac-de-mot) • Ignorent l'aspect sémantique
	Méthodes basées sur les graphes	• Modélisation sémantique, contextuelle et structurelle	• Graphes complexes dans le cas des textes longs • Temps de calcul élevé
Approche sémantique		• Représentation conceptuelle et sémantique riche • Espace de représentation réduit • Vocabulaire contrôlé	• Limitées au domaine décrit par la ressource sémantique utilisée • Pour la langue arabe, les ressources comme les ontologies et les thésaurus standardisés sont rares
Approches hybrides		• Compromis entre différentes approches	• Accroissement de la complexité par rapport aux systèmes classiques

Tableau 5 : Synthèse des approches et méthodes d'indexation

5. Conclusion

Dans cet article, nous avons présenté différentes techniques d'indexation automatique. Le choix d'une bonne méthode de représentation du document influencera significativement les étapes ultérieures d'analyse de documents. Cependant, plusieurs critères sont mis en question, notamment la réduction de dimension, la conservation du contexte et de la sémantique.

Les auteurs favorisent les méthodes statistiques pour la représentation des documents arabes, considérant la simplicité de leur traitement. Nous recensons peu de travaux sur les documents arabes qui s'intéressent à exploiter de nouvelles méthodes d'extraction de descripteurs, ceci est dû essentiellement à la complexité de la structure de cette langue.

Comme perspective à cette contribution, nous proposerons une nouvelle approche orientée contexte qui tire profit des méthodes classiques, tout en palliant à certaines limites de l'existant. Dans nos futurs travaux, nous souhaitons accorder plus d'importance à la phase d'indexation, et ce, en essayant d'améliorer les méthodes déjà existantes et tenter des hybridations entre différentes techniques. Notre objectif est de proposer un système d'indexation faisant face aux trois enjeux suivants : la sémantique, l'espace de représentation et la complexité de calcul.

6. Références

Al Molijy, A., Hmeidi, I. & Alsmadi, I. (2012) *Indexing of Arabic documents automatically based on lexical analysis*. International Journal on Natural Language Computing (IJNLC) Vol. 1, No.1.

Alajmi, A., Saad, E.M., Darwish, R.R. (2012) *Toward an ARABIC Stop-Words List Generation*. International Journal of Computer Applications (0975-8887) Volume 46- No.8.

Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M. S. & Al-Rajeh, A. (2008). *Automatic Arabic Text Classification*. In Proceedings of The 9th International Conference on the Statistical Analysis of Textual Data, JADT.

Al-Mahmoud, H., Al-Razgan, M. (2015). *Arabic Text Mining: A Systematic Review of the Published Literature 2002-2014*, International Conference on Cloud Computing (ICCC).

Alsaleem, S. (2011). *Automated Arabic Text Categorization Using SVM and NB*. International Arab Journal of e-Technology, vol. 2, no. 2.

Al-Salemi, B. & Aziz, M. J. A. (2011). *Statistical Bayesian Learning For Automatic Arabic Text Categorization*. Journal of Computer Science, vol. 7, no. 1, pages 39–45.

Al-Shalabi, R. & Obeidat, R. (2008). *Improving KNN Arabic Text Classification with N-Grams Based Document Indexing*. In Proceedings of the Sixth International Conference on Informatics and Systems, INFOS, pages 108–112.

Al-Shalabi, R. Kanaan, G. & Gharaibeh, M. (2006). *Arabic Text Categorization Using kNN Algorithm*. In Proceedings of The 4th International Multiconference on Computer Science and Information Technology, volume 4 of CSIT'2006.

Al-Shalabi, R. Kanaan, G. & Gharaibeh, M. (2010). *Arabic Text Categorization Using kNN Algorithm.* In Proceedings of the 6th International Conference on Advanced Information Management and Service, IMS. Institute of Electrical and Electronics Engineers (IEEE).

Alwedyan, J. Musa, W. H., Salam, M. & Mansour, H. Y. (2011). *Categorize arabic data sets using multi-class classification based on association rule approach*. In Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, ISWSA'11, New York, NY, USA, ACM, pages 18 :1–18 :8.

Bawaneh, M. J. Alkoffash, M. S. & Al Rabea, A. I.. (2008). *Arabic Text Classification using K-NN and Naive Bayes*. Journal of Computer Science, vol. 4, pages 600–605.

Benkhalifa, M. Mouradi, A. & Bouyakhf, H. (2001). *Integrating WordNet knowledge to supplement training data in semi-supervised agglomerative hierarchical clustering for text categorization*. International Journal of Intelligent Systems, vol. 16, no. 8, pages 929–947.

Bessou, S., Saadi, A. et Touahria, M. (2007). *Un système d'indexation et de recherche des textes en arabe (SITRA).* 1er séminaire national sur le langage naturel et l'intelligence artificielle (LANIA), Université HAssiba ben Bouali, Département d'Informatique, Chlef (DZ), 20-21.

Bougouin, A. (2013). *État de l'art des méthodes d'extraction automatique de termes-clés.* TALN-RÉCITAL 2013, 17-21 Juin, Les Sables d'Olonne.

Chantar, H. K. & Corne D. W. (2012). *Arabic Text Categorization via Binary Particle Swarm Optimization and Support Vector Machines.* In The 5th International Conference on Bioinspired Optimization Methods and their Applications, BIOMA'2012.

Duwairi, R. M. (2005). *A Distance-based Classifier for Arabic Text Categorization.* In In Proceedings of The 2005 International Conference on Data Mining, DMIN'2005, CSREA Press, pages 187–192.

Duwairi, R. M. (2006). *Machine learning for Arabic text categorization: Research Articles*. Journal of American society for Information Science and Technology, vol. 57, no. 8, pages 1005–1010.

El Kourdi, M., Bensaid, A. & Rachidi, T. (2006). *Automatic Arabic document categorization based on the Naive Bayes algorithm*. In Proceedings of the Workshop on Computational Approaches to Arabic Script based Languages, SEMITIC '04, Stroudsburg, PA, USA. Association for Computational Linguistics, pages 51–58.

El-Halees, A. M. (2007). *Arabic Text Classification Using Maximum Entropy.* The Islamic University Journal (Series of Natural Studies and Engineering), vol. 15, no. 1, pages 157–167.

El-Khoribi, R. A. and. Ismael, M. A (2006). *An Intelligent System Based on Statistical Learning For Searching in Arabic Text.* ICGST International Journal on Artificial Intelligence and Machine Learning, AIML, vol. 6, pages 41–47.

El-Kourdi, M. Bensaid, A. & Rachidi, T. (2010). A*utomatic Arabic Document Categorization Based on the Naïve Bayes Algorithm*. In Proceedings of The 7th International Conference on Informatics and Systems, INFOS.

Gharib, T. F. Habib, M. B. & Fayed, Z. T. (2009). *Arabic Text Classification Using Support Vector Machines*. International Journal of Computers and Their Applications ISCA, vol. 16, no. 4, pages 192–199.

Harrag, F., El-Qawasmah, E. & Al-Salman, A. (2011). *Stemming as a Feature Reduction Technique for Arabic Text Categorization*. In Proceedings of The 10th International Symposium on Programming and Systems, ISPS, pages 128–133.

Hasan, K.S. & Ng, V. (2010). *Conundrums in unsupervised keyphrases extraction : Making sense of the state of the art.* Proceedings of the 23rd International Conference on Computatioonal Linguistics (COLING-10), Poster Volume.

Herskovic, J. R. & Jorge, R. (2011). *MEDRank: Using graph-based concept ranking to index biomedical texts.* International Journal of Medical Informatics volume 80 issue 6 pages : 431-441.

Hmeidi, I., Hawashin, B., El-Qawasmeh, E. (2008). *Performance of KNN and SVM classifiers on full word Arabic articles.* Advanced Engineering Informatics 22, pages 106–111.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval Pages 50-57.

Hulth, A. (2003). *Improved automatic keyword ex- traction given more linguistic knowledge.* In Proceedings of EMNLP, pages 216–223.

Jamoussi, S. (2009). *Une nouvelle représentation vectorielle pour la classification sémantique*. TAL volume 50.

Joachims, T. (1998). *Text Categorization with Suport Vector Machines : Learning with Many Relevant Features*. In Proceedings of the 10th European Conference on Machine Learning, ECML'98, London, UK, UK. Springer-Verlag, pages 137–142.

Kanaan, G., Yaseen, M., Al-Shalabi, R., Al-Sarayreh, B. & Mustafa, A.. (2009). *Using EM for Text Classification on Arabic.* In Proceedings of the Second International Conference on Arabic Language Resources and Tools. The MEDAR Consortium.

Kanaan, G., Al-Shalabi, R. & AL-Akhras, A. (2006). *KNN Arabic Text Categorization Using IG Feature Selection.* In Proceedings of The 4th International Multiconference on Computer Science and Information Technology, volume 4 of CSIT'2006.

Khreisat, L. (2006). *Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study*. In Proceedings of The 2006 International Conference on Data Mining, DMIN '2006, CSREA Press, pages 78–82.

Li, H. Y. & Jain, K. A. (1998). *Classification of text documents*. The Computer Journal, vol. 41, no. 8, pages 537–546.

Liu, Z. , Peng, L. , Yabin, Z. & Maosong, S. (2009). *Clustering to find exemplar terms for keyphrase extraction*. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 257–266.

Mallak, I. (2011). *De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information.* Thèse pour l'obtention du grade de docteur. Université Paul Sabatier – Toulouse III. France.

Mansour, N., Haraty, R.A., Daher, W., Houri, M. (2008). *An auto-indexing method for Arabic text*. Information Processing and Management, volume: 44 issue: 4, pages: 1538-154.

Matsuo, Y et Ishizuka, M. (2004). *Keyword Extraction From a Single Document Using Word Co-Occurrence Statistical Information*. International Journal on Artificial Intelligence Tools volume 13 issue 1 pages: 157-169.

Mesleh, A. M. & Kanaan, G. (2008). *Support vector machine text classification system: Using Ant Colony Optimization based feature subset selection*. In proceeding of the International Conference on Computer Engineering & Systems, ICCES '2008, pages 143–148.

Mesleh, A. M. (2007b). *CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System.* Journal of Computer Science, vol. 3, no. 6, pages 430–435.

Mesleh, A. M. (2007c). *CHI Square Feature Extraction Based SVMs Arabic Language Text Categorization System.* In proceeding of the 2nd International Conference on Software and Data Technologies, (Knowledge Engineering), pages 235–240.

Mesleh, A. (2007a). *Support vector machines based Arabic language text classification system : feature selection comparative study.* In Proceedings of the 12th WSEAS International Conference on Applied Mathematics, MATHq07, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society (WSEAS), pages 11–16.

Mihalcea, R. & Tarau, P. (2004). *Textrank: Bring- ing order into texts*. In Proceedings of EMNLP, pages 404–411.

Mohamed, R. & Watada, J. (2010). *An Evidential Reasoning Based LSA Approach to Document Classification for Knowledge Acquisition.* In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management, IEEM'10. Institute of Electrical and Electronics Engineers (IEEE), pages 1092–1096.

Motasem, A. & Joseph, D. (2009). *Levée d'ambigüité par la méthode d'exploration contextuelle: la séquence 'alif-nûn* (ان) *en arabe*. In Ghenima, Malek, Ouksel, Aris et Sidhom, Sahbi (eds.), Systèmes d'Information et Intelligence Economique, 2ème Conférence Internationale, organisée par l'université de Nancy, France et l'université de la Manouba, École supérieure de commerce électronique (ESCE), Tunis, Tunis, Hammamet, IHE éditions, pages. 573-585.

Mountassir, A. (2012). *Sentiment Analysis: Classification supervisée de documents arabes.* Proceedings of 7th International Conference on Intelligent Systems : Theories and Applications. Mohammedia, Morocco.

Page, L. Brin, L. Motwanin R., & Winograd, T. (1998). *The pagerank citation ranking: Bringing order to the web.* Technical report, Stanford Digital Library Technologies Project, 1998.

Porter, M.F. (1980). *An algorithm for suffix stripping*. Program, Vol. 14 No.3, pp. 130-137.

Raheel, S. & Dichy, J. (2010). *An empirical study on the feature qs type effect on the automatic classification of arabic documents.* In Proceedings of the 11th international conference on Computational Linguistics and Intelligent Text Processing, CICLing'10 Berlin, Heidelberg, pages 673–686.

Raheel, S., Dichy, J. & Hassoun, M. (2009). *The Automatic Categorization of Arabic Documents by Boosting Decision Trees*. In Proceedings of the Fifth International Conference on Signal Image Technology and Internet Based Systems Washington, DC, USA. IEEE Computer Society, pages 294–301.

Robertson, S.E., Walker, S. & Beaulieu, M. (2000). *Experimentation as a way of life: Okapi at TREC.* Information Processing and Management, vol. 36, pages 95–108.

Roche, M. (2011). *Fouille de Textes : De l'extraction des descripteurs linguistiques à leur induction.* Thèse, université Montpellier II, France.

Saadane, H. (2013). *Une approche linguistique pour l'extraction des connaissances dans un texte arabe.* TALN-Récital, 17-21 juin, Les Sables d'Olonne.

Salton, G., Wong, A. & Yang, C. S. (1975). *A vector space model for automatic indexing. Commun*. ACM, vol. 18, no. 11, pages 613–620.

Samir, A. M., Ata, W. & Darwish, N. (2005). *A New Technique for Automatic Text categorization for Arabic Documents*. In Proceedings of the 5th Conference of the Internet and Information Technology in Modern Organizations, pages 13–15.

Sawaf, H., Zaplo, J. & Ney, H. (2001). *Statistical Classification Methods for Arabic News Articles*. In Arabic Natural Language Processing Workshop, ACL'2001, Retrieved from Arabic NLP Workshop at ACL/EACL 2001 website : `http ://www.elsnet.org/acl2001-arabic.html`, pages 78–82.

Schapire, R. E. & Singer, Y. (2000). *BoosTexter: A Boosting-based System for Text Categorization*. Machine Learning, vol. 39, no. 2/3, pages 135–168.

Schapire, R. E. Singer, Y. & Singhal, A. (1998). *Boosting and Rocchio applied to text filtering.* In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '98, New York, NY, USA. ACM, pages 215–223.

Shafiei, M., Wang, S., Zhang, R., Milios, E., Tang, B., Tougas, J. & Spiteri, R. (2007). *Document Representation and Dimension Reduction for Text Clustering.* In Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop, ICDEW q07, Washington, DC, USA. IEEE Computer Society, pages 770–779.

Stetina J., Kurohashi S. et Nagao. M. (1998). General word sense disambiguation method based on a full sentential context. Usage of WordNet in Natural Language Processing, Proceedings of COLING-ACL Workshop, Montreal, Canada, July 1998

Syiam, M. M., Fayed, Z. T. & Habib, M. B. (2006). *An Intelligent System For Arabic Text Categorization.* International Journal of Intelligent Computing and Information Sciences, vol. 6, no. 1, pages 1–19.

Thabtah, F., Eljinini, M., Zamzeer, M. & Hadi, W. (2009). *Naïve Bayesian based on Chi Square to Categorize Arabic Data*. In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, IBIMA'2009, pages 930–935.

Thabtah, F., Hadi, W. & Al-shammare, G. (2008). *VSMs with K-Nearest Neighbour to Categorise Arabic Text Data*. In Proceedings of The World Congress on Engineering and Computer Science, WCECS '2008, pages 778–781.

Tomokiyo T. et Hurst M.. (2003). A language model approach to keyphrase extraction. In Proceedings of the ACL Workshop on Multiword Expressions.

Trstenjak, B., Mikac, S., Donko, D. (2013). *KNN with TF-IDF Based Framework for Text Categorization*. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation.

Wan, W et Xiao, J. (2008a). Collabrank: Towards a collaborative approach to single-document keyphrase extraction. In Proceedings of COLING, pages 969–976.

Wan, X et Xiao, J. (2008b). *Single document keyphrase extraction using neighborhood knowledge*. In Proceedings of AAAI, pages 855–860.

Wei, G., Gao, X.and Wu, S. (2010). *Study of Text Classification Methods for Data Sets With Huge Features*. In Proceedings of the 2nd International Conference on Industrial and Information Systems, volume 1, pages 433–436.

Wei,W. (2013). Regroupement sémantique de relations pour l'extraction d'information non supervisée. TALN-RÉCITAL, Les Sables d'Olonne.

Yang, Y. & Chute, G. C. (1994). *An example-based mapping method for text categorization and retrieval.* ACM Transactions on Information Systems, vol. 12, no. 3, pages 252–277.

Yongjing, L. (2007). *A Document Clustering and Ranking System for Exploring MEDLINE Citations*. Journal of the American Medical Informatics Association volume 14 issue 5, pages: 651-661.

Zaki, T. (2013). *Indexation par le contenu et archivage de fonds documentaires arabes*. Thèse pour l'obtention du grade de doctorat d'université, Université Ibn Zohr, Agadir, Maroc.

Zaki, T. Mammass, D. Ennaji A., Nicolas, S. (2014). *A kernel hybridization NGram-Okapi for indexing and classification of Arabic documents*. Journal of information and Computing Science. ISSN 1746-7659, England, UK. Vol. 9 No.2. pages:141-153.

Zaki, T. Mammass, D. Ennaji, A. (2010). *A semantic proximity based system of Arabic text indexation*. International Conference on Image and Signal Processing (ICISP).

Zargayouna, H. et Salotti, S. (2004). *Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML*. IC: 15es journées francophones d'ingénierie des connaissances.

Zubi, Z. S. (2009). *Using some web content mining techniques for Arabic text classification*. In Proceedings of the 8th WSEAS international conference on Data networks, communications, computers, Stevens Point, Wisconsin, USA. World Scientific and Engineering Academy and Society, pages 73–84.