Le Text Mining sur la langue Arabe : application au traitement des sources ouvertes[1]

Charles Huot (*) et Pascal Coupet (*)
Charles.Huot@temis-group.com, Pascal.Coupet@temis-group.com

(*) : TEMIS SA, Tour Gamma, 193-197 rue de Bercy, 75582 Paris cedex 12, www.temis.com .

Mots clés : text mining, analyse automatique de textes, langue Arabe, sources ouvertes, contre terrorisme.

Résumé

Les événements dramatiques du 11 septembre 2001 ont forcé les gouvernements et notamment le gouvernement Américain à renforcer ses investissements dans les logiciels de traitement automatique du langage à des fins d'intelligence et de contre terrorisme.

Le Text Mining fait appel à diverses méthodes d'analyse, comme la linguistique, la classification automatique ou la catégorisation. L'application de ces méthodes, nécessite en fonction du type d'indicateur que l'on souhaite mettre en place, une plus ou moins grande connaissance formalisée du domaine couvert par les documents à analyser.

Nous présenterons dans cet article, la combinaison de 2 méthodes, l'analyse linguistique de l'Arabe et la classification automatique. L'objectif est de mettre en place une chaîne de traitement capable d'analyser de manière automatique des textes en langue Arabe en provenance de sources ouvertes.

[1] Marques déposées : Insight Discoverer, Cartouches de connaissances, Skill Cartridges et Online Miner et Xelda sont des marques déposées par TEMIS SA en France et dans le reste du monde.

1. Introduction

Il est des périodes où certains développements techniques sont favorisés par le contexte politique du moment. C'est le cas aujourd'hui pour toutes les méthodes et techniques de capture, de saisie, de reconnaissance, de traduction et d'analyse de la langue arabe. Un grand nombre d'éditeurs de logiciels, notamment américains, font avancer le sujet sur l'ensemble de ces fronts. Les utilisateurs de ces logiciels sont pour la grande majorité, les services des états, et notamment les services de police et de renseignement.

Mises à part les entreprises qui fabriquent des moteurs de recherche (Yahoo!, Google,…) peu d'applications civiles en font émerger le besoin.

Nous présenterons dans cet article une articulation entre une approche traditionnelle linguistique de l'analyse de documents arabes et d'un système de classification automatique (clustering) de cartographie de documents. Afin de présenter nos travaux dans un contexte plus général des applications de l'analyse automatique de la langue arabe nous introduirons notre propos par un certain nombre de réalisations existantes sur le sujet.

2. Les applications de l'analyse linguistique Arabe

Les américains ont développé un ensemble d'applications. Une présentation récente (avril 2005, Phoenix, USA) de la société *In-Q-Tel* nous a servi de base pour en décrire un certain nombre. [1]

Operation Just Cause (12/1989)
- 6 million documents
- 100 tonnes
- 0.5 mile stack

Operation Desert Storm (1/1991)
- 12 million documents
- 200 tonnes
- 1.1 mile stack

Operation Iraqi Freedom (2/2004)
- 100 million documents
- 2,000 tonnes
- 12 mile stack

Le tableau ci-dessus nous indique l'augmentation du volume de documents relatifs aux opérations ***Juste Cause*** (Décembre 1989), ***Desert Storm*** (Janvier 1991) et ***Iraqi Freedom*** (Février 2004). L'indication est donnée en nombre de documents, poids et miles linéaires. Entre la première opération et la dernière, les indicateurs ont été multipliés par un facteur de l'ordre de 15 à 20.

Les efforts dans le traitement automatisé de ce type de volume d'information ont porté sur quatre aspects :

- la reconnaissance optique des caractères,
- la traduction automatique,
- le traitement de la voix,
- la recherche en langue anglaise sur des documents en arabe.

2.1. La reconnaissance optique de caractères

Les méthodes de traitement de l'information nécessitent une forme électronique des documents. Pour accéder à la compréhension du document lui-même, cette forme ne doit pas être une simple image. Il est nécessaire de mettre en place une chaîne de reconnaissance optique des caractères sur les

images issues de la phase de scanner des documents originaux. Outre la difficulté inhérente à l'écriture elle-même il est nécessaire de rectifier certains points liés à la mauvaise qualité des documents papiers d'origine (photocopie, fax, etc…).

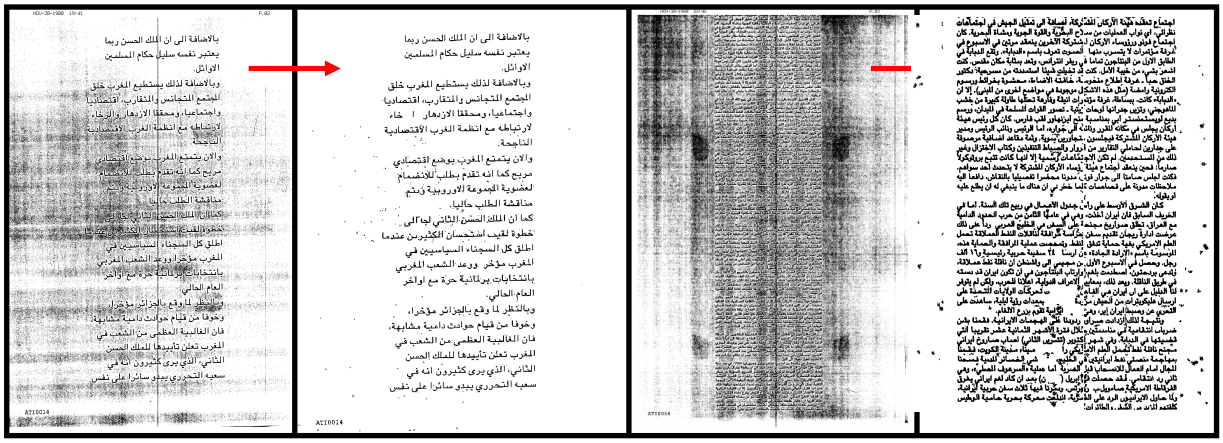
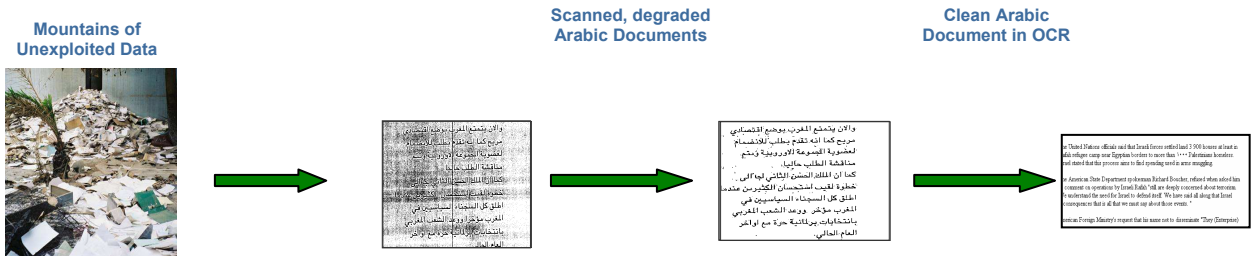


Figure 1 : Nous pouvons voir sur cette figure le travail de nettoyage réalisé par le logiciel avant le lancement de la phase de reconnaissance optique de caractère. (Source : *NovoDynamics Inc.*)

Nous sommes dans un schéma comprenant ici quatre phases.

-Phase 1 : Recueil des documents sur le terrain.
-Phase 2 : Scanner des documents
-Phase 3 : Redressement et nettoyage des résultats du scanner (figure 1)
-Phase 4 : Reconnaissance optique des caractères pour création d'un fichier de texte électronique.



2.2. La traduction automatique

La traduction automatique est également un grand champ d'application de la linguistique Arabe. Le développement du web contribuant largement à la diffusion des sites en langue arabe. L'exemple ci-dessous (figure 2) montre une application de traduction instantanée de l'arabe vers l'anglais.

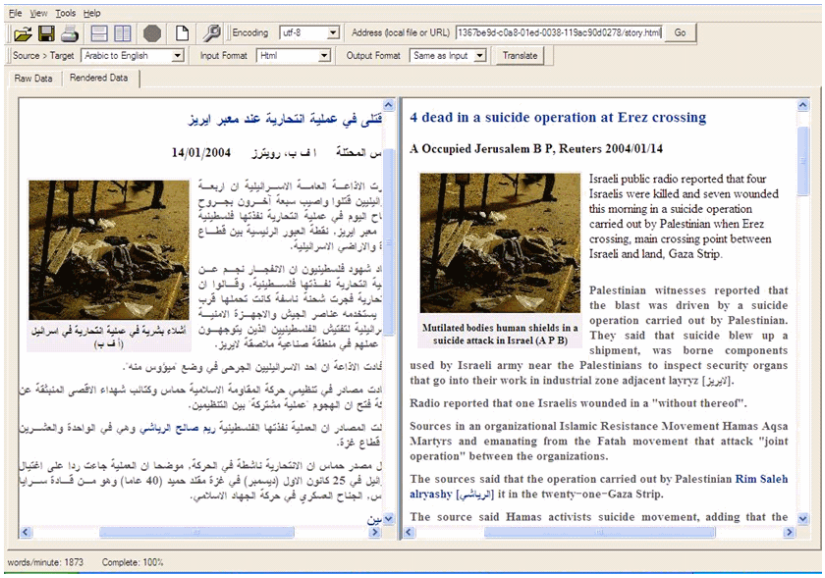
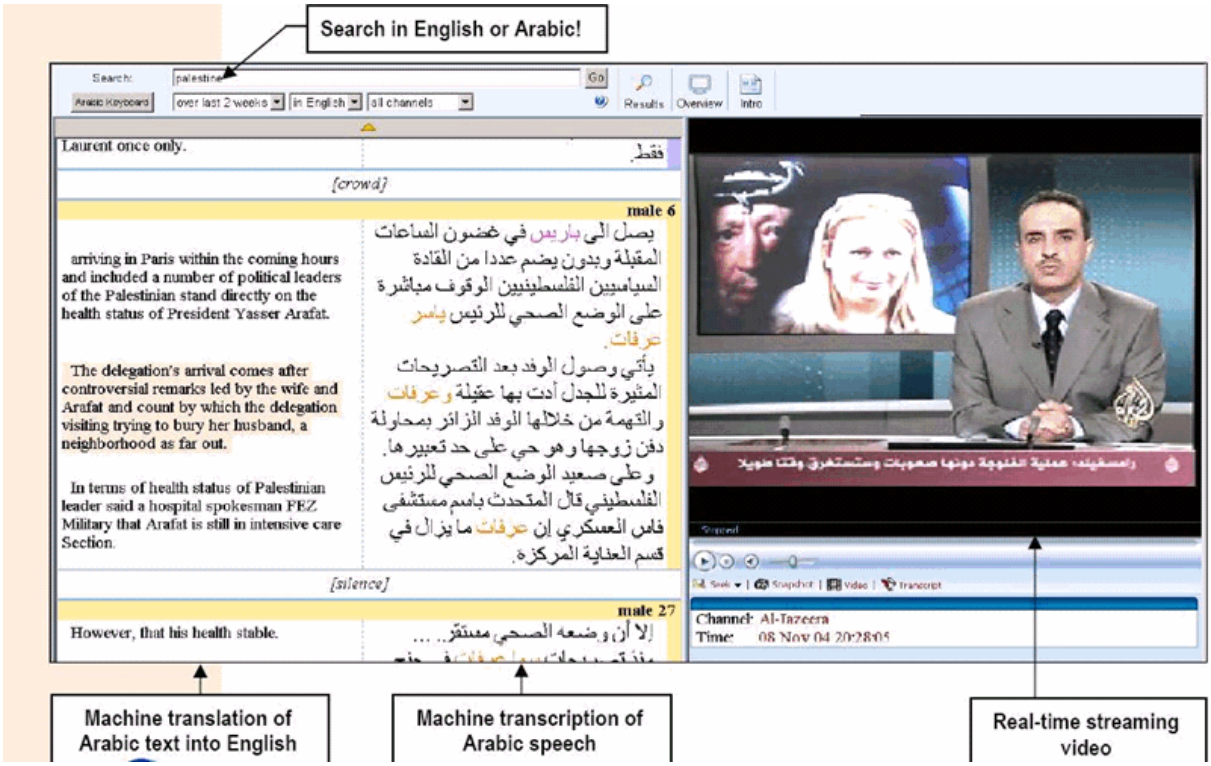


Figure 2 : Source : *Language Weaver*

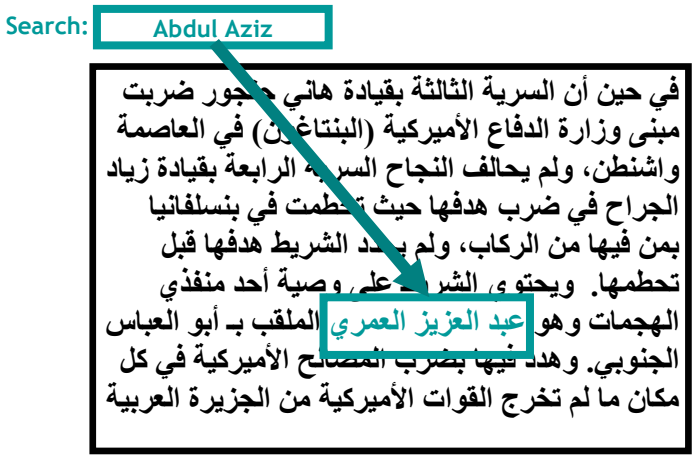
2.3. Le traitement automatisé de flux vidéo en langue arabe



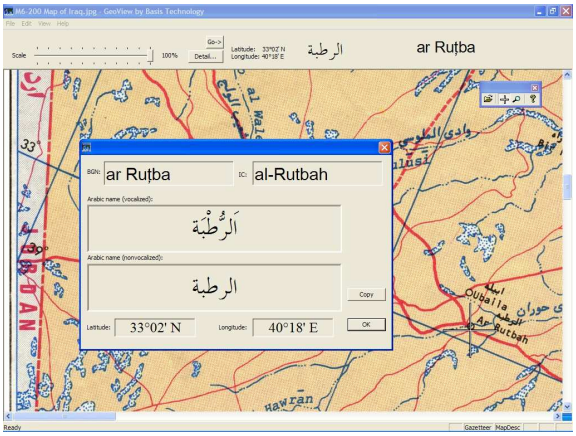
Source : *In-Q-Tel, BBN*

Cet exemple nous montre une combinaison de différentes technologies et, notamment de retranscription du commentaire du journaliste, d'une forme verbale vers une forme écrite (Machine transcription of Arabic speech) avec en seconde étape une traduction instantanée de la forme écrite de l'Arabe vers l'Anglais.

2.4. La recherche en anglais sur des documents en langue arabe



Exemple de recherche en anglais sur un texte en langue arabe



Exemple de recherche en anglais sur la base d'un rapprochement phonétique et transcription en arabe pour réaliser la recherche sur une carte géographique.

Source : *Basis Technology Inc.*

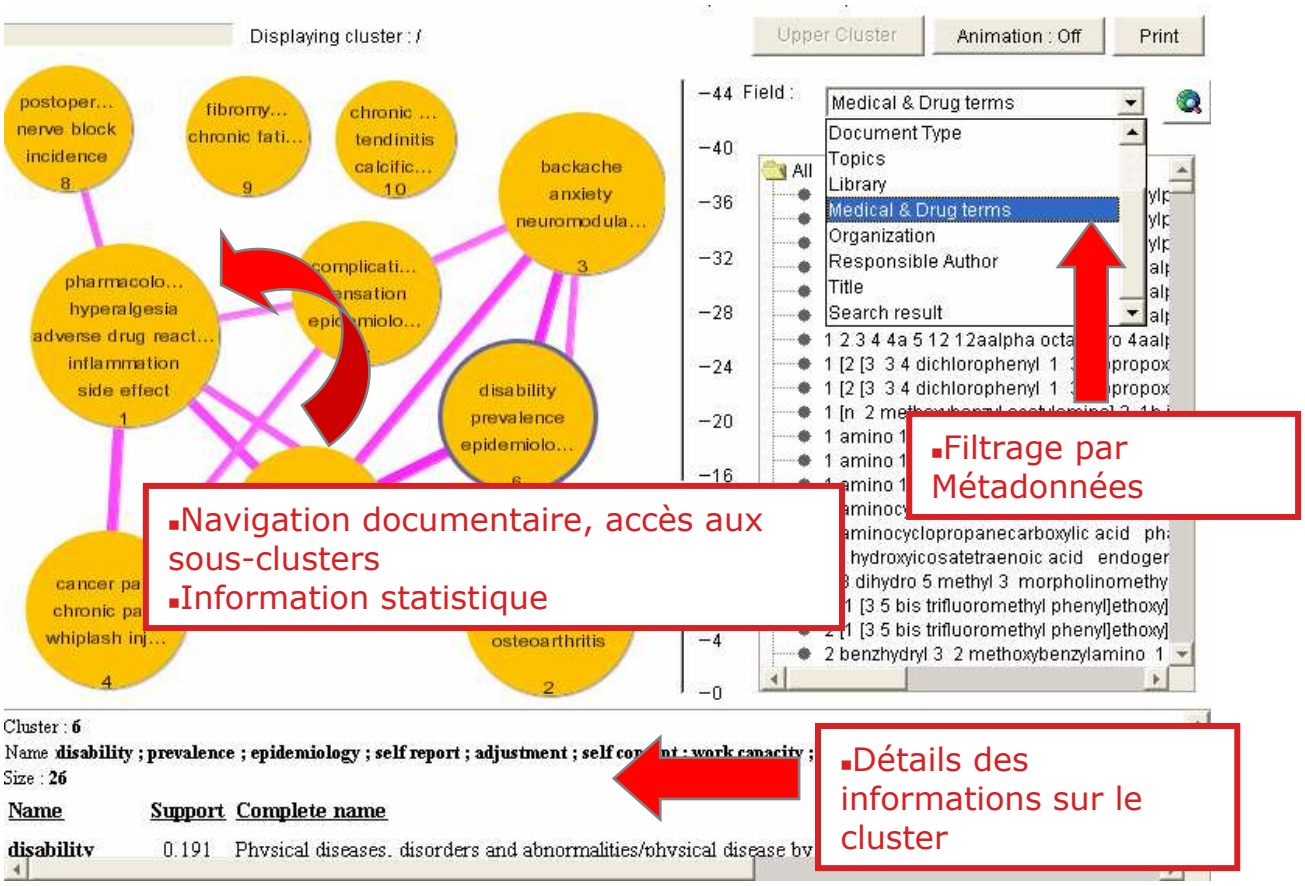
3. La classification automatique (Clustering) de documents en langue Arabe

3.1. La classification comme un outil d'analyse de documents

De très nombreux travaux dans notre discipline ont démontré les bénéfices de l'application des méthodes de classification automatique dans l'analyse de grands volumes de documents. L'équipe Centre Européen de Mathématique Appliqués d'IBM a développé, dès les années 1990, sur la base de la méthode de l'Analyse Relationnelle des données, inventée par Jean-François Marcotorchino et Pierre Michaud [2], des logiciels très performants pour des applications dans le domaine de la veille technologique et de l'intelligence économique [3]. L'équipe Toulousaine du Professeur Dousset (IRIT) sur la base de méthodes de classification hiérarchique, a travaillé depuis des années sur des logiciels pour des entreprises et des instituts

de recherche dans ce même domaine [4]. Enfin les équipes du professeur Henri Dou à Marseille, au CRRM (Centre de Recherche Rétrospective de Marseille) ont également largement supporté ce champs de recherche et de développement de logiciels adaptés[5], [6], [7].

Plus récemment, l'équipe de TEMIS, à travers son logiciel Insight Discoverer Clusterer et son application Online Miner, soutient les grandes entreprises européennes en complétant leurs applications de gestion électronique de documents ou leur moteur de recherche avec une fonctionnalité de clustering. Ceci permet à un utilisateur de cartographier automatiquement une liste de plusieurs centaines ou milliers de documents, issus d'une requête, et de faciliter sa navigation en découvrant automatiquement les thèmes et sous thèmes, émergeant des documents.



3.2. Présentation de la méthodologie

3.2.1. L'analyse linguistique

TEMIS a intégré dans sa chaîne de traitement automatique des langues, un analyseur morphosyntaxique de la langue arabe.

"The general rule of thumb for Arabic is that *everything is at least five times more complicated than for any European language."* ***(Ken Beesley)***

Même si il existe des mots Arabe s'appuyant sur des stems qui ne peuvent pas être décomposés en racine et affixes (préfixe et suffixe) (c'est le cas notamment pour les mots étrangers), la plupart des noms et des verbes possèdent une racine et des dérivations.

Ces racines véhiculent une idée principale. Par exemple, *k-t-b* indique l'idée d'écriture. Lorsque l'on accole d'autres lettres, avant, entre ou après les lettres de base, cela nous donne un ensemble de mots associés : « écrire », mais aussi « livre », « bureau », « bibliothèque » et « auteurs ». (http://www.al-bab.com/arab/language/lang.htm).

Par exemple la racine correspondant au concept de d'apprentissage et de formation est *d-r-s*. Même si la forme de base reste *drs*, le schéma et la vocalisation vont varier selon l'usage:

darasa, "étudier, apprendre"
darrasa, "enseigner"
durus, "leçons"
mudaaris, "enseignant"/mudaarisa, "enseignante"
madrasa, "école"
http://saturn.sron.nl/~jheise/akkadian/semitic.html

Les ambiguïtés, dans un texte, sont ainsi résolues généralement lorsque l'on est dans une forme vocalisée.

Les voyelles sont notées en utilisant les marques diacritiques. Les voyelles longues sont toujours écrites, même dans les formes non vocalisées, alors que les voyelles courtes sont généralement supprimées. Ainsi la racine ne correspond pas toujours exactement à la forme non vocalisée.
Par exemple, « ktb » est la racine de livre « kitAb », mais la forme non vocalisée du mot est « ktAb » (le 'A' étant le début d'une longue voyelle ')

L'un des problèmes également rencontré, est que la majorité des documents arabes sont écrits sans les voyelles.

3.2.2. La classification automatique

Dans le cadre de cette recherche, TEMIS a utilisé son logiciel standard Insight Discoverer Clusterer. Il s'agit d'un serveur de classification automatisée qui regroupe dynamiquement les documents en fonction de leur ressemblance sémantique et de leur proximité thématique. Il propose le classement le plus pertinent pour un fond documentaire donné. Les utilisateurs peuvent ainsi naviguer dans leurs documents organisés par thèmes et sous thèmes. Ils disposent à la fois d'une vue d'ensemble de l'information et de différentes pistes d'exploration.

L'analyse morphosyntaxique en Arabe, réalisée par notre logiciel d'extraction Insight Discoverer Extractor, permet de générer des descripteurs en Arabe pour chaque document. Ensuite l'algorithme de clustering regroupe les documents similaires dans des classes. Cet algorithme est spécialement adapté à l'analyse textuelle. Il est possible de paramétrer la profondeur du plan de classement et le nombre de classes souhaitées par niveau. Un titre est attribué à chaque classe qui reprend, par ordre hiérarchique, les termes ou expressions les plus caractéristiques de la classe.

3.2.3. La traduction automatique des résultats

Afin de faciliter l'analyse du résultat d'une classification de documents un module de traduction automatique des termes qui composent les titres de chaque classe est mis en place. Ce module (MouseOver™), développé par la société Basis Technology (Boston, USA), s'appuie sur des dictionnaires bilingues qui remplacent mot à mot, les termes figurant dans les titres.

Cette approche a l'avantage de bien expliciter le contenu d'un cluster sans avoir à faire la traduction de tous les documents. Cela facilite le travail de navigation et de découverte du corpus de texte à analyser. Il permet une plus grande autonomie de l'utilisateur.

3.3. Exemple de résultats

Afin de valider notre approche, nous avons réalisé une série d'études sur des documents arabes en provenance de différentes sources. L'Agence France Presse est l'un des plus gros diffuseurs européen de dépêches en langues Arabe. En accord avec eux, nous avons utilisé leur source comme corpus d'évaluation. La copie d'écran ci-dessous présente un exemple de rapport de clustering.

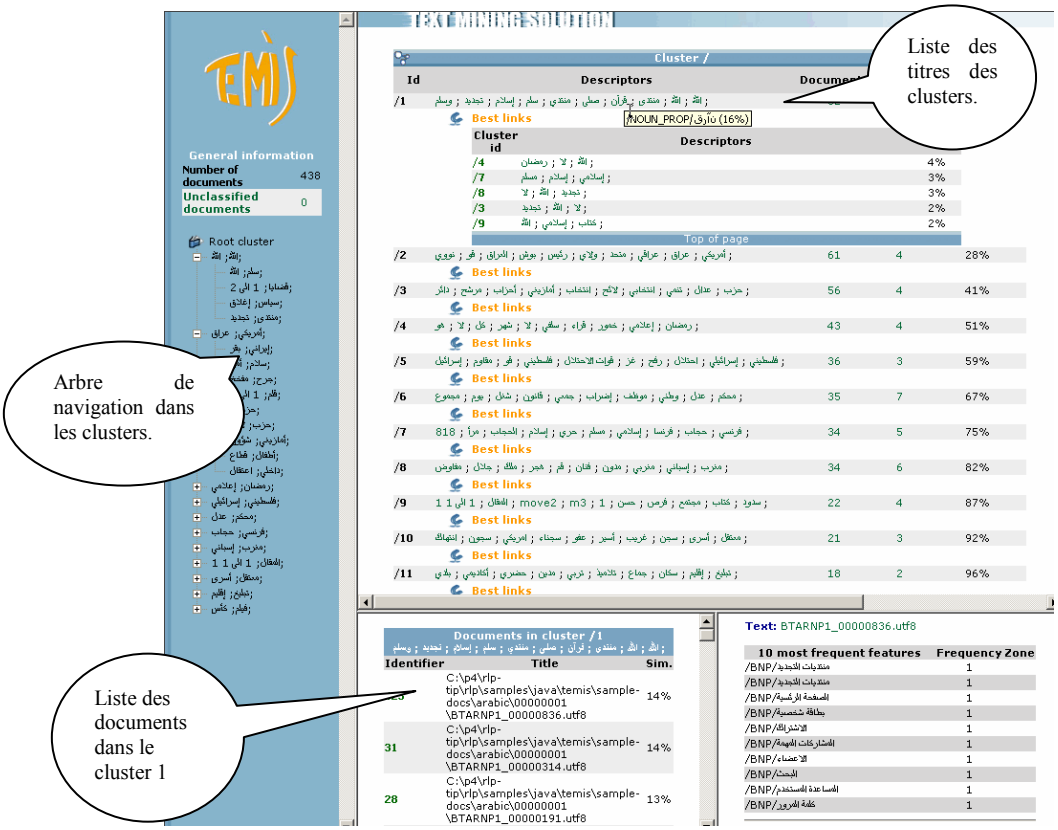


Figure 3 : Interface de navigation du rapport de clustering

Ce rapport permet à l'utilisateur de naviguer de manière intelligente au sein de l'ensemble des documents de l'étude.

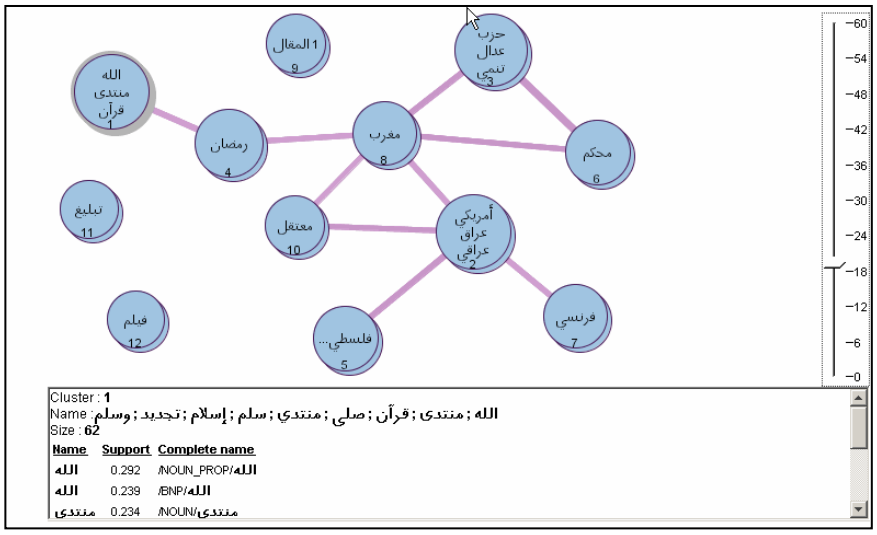


Figure 4 : ce graphe de navigation généré au moment de la création du rapport est un autre moyen de ce déplacer dans les résultats de l'étude. Il indique notamment les relations principales qui existent entre les différents thèmes identifiés par le clustering.

4. Conclusion et perspectives

Cette nouvelle approche est intéressante et très utile, compte tenu de l'augmentation des volumes d'information à analyser en langue arabe. Elle permet à l'utilisateur de se faire rapidement une opinion sur le contenu d'un ensemble de documents écrits dans cette langue afin de réaliser des tris ou des filtrages, selon des critères qu'il aura déterminés.

Aujourd'hui nous poursuivons nos travaux sur le développement de cartouches de connaissances en langue arabe à des fins d'extraction automatique de concepts et de relations spécifiquement recherchées. [8]. Les premières études de faisabilité ont montré une compatibilité totale avec les logiciels de TEMIS.

D'autres travaux sont également en cours sur l'application de ces mêmes méthodes sur l'analyse de sources ouvertes en langues complexes comme le chinois, le japonais ou le coréen.

5. Bibliographie

[1] Greg Pepus, In-Q-Tel, Inc.(2005), Information Intelligence Summit, April 20th, 2005 Delphi Group, Phoenix, USA

[2] François Marcotorchino et Piierre Michaud, “optimisation en analyse ordinale des données”, 210p, Masson, Paris, 1979.

[3] Chantal Béfécarrax et Charles Huot, Application de l’analyse relationelle à la veille technologique: des outils d’analyse de l’information documentaire, Revue française de bibliométrie, Vol 9, p66-80, Sept 1991.

[4] Saïd Karonach et Bernard Dousset, “Visualisation interactive pour la découverte de connaissances: GeoECD, Congrès VSST 2001, 15-19 Octobre 2001. Barcelone, Espagne.

[5] Rostaing H. , "La bibliométrie et ses techniques." , Sciences de la Société , Collection , (1996)

[6] Rostaing H., Nivol W., Quoniam L., Bédécarrax C., Huot C. , "L'exploitation systématique des bases de données : des analyses stratégiques pour l'entreprise" , Les cahiers de l' ADEST , Juillet, p. 7-22 , (1993)

[7] La veille technologique, sous la direction de H.Dou et H.Desval, Dunod, 1992.

[8] Pascal Coupet, Bianka Buschbeck, Amandine Six, Françoise Cardoso et Charles Huot, "Le Text Mining multilingue: application au monde de l’intelligence économique", Congrès SFBA 2005, 13-17 Juin 2005. Ile Rousse, France.