"Al-Hadiths" Information Retrieval

Fouzi HARRAG
Université de Sétif
Aboubekeur HAMDI-CHERIF
Computer College, Qassim University
Buraydah, Saudi Arabia

Abstract*:*
We are interested in text mining as applied to the Arabic language using the vector space research model. We show that the use of Arabic roots as means of indexing terms as well as the use of statistical methods has a direct impact on the research performance within the vector space model. The proposed system is a text mining called **Authentique** (in French language **Authen**tification automa**tique**) whose aim is to provide a list of prophetic traditions "Hadiths" classified according to their degrees of similarity based on a given query. Among the other implemented classic text mining methods in Authentique are TFIDF Weight and cosine measure.

1. Introduction

Textual information takes more and more importance in the daily activity of researchers and companies, but the quantity of this information brings about problems of access and search, (Bilhaut, 2006). In this article, we adapt techniques based on a vector space model indexing. It is about measuring, thanks to information recorded in an index, the semantic similarity, or score, between every text of the database and the user's query by means of a similarity measure. To finish, the documents found are ordered according to their score and then proposed to the user. In order to get full lists of good quality documents, we chose to use a proven method, *i.e.* the so-called vector space model as a basis for the proposed research system, called *Authentique*, (Harrag & Hamdi-Cherif, 2006). This model was notably criticized because of the hypothesis of independence of the keyword (the dimension of the space corresponds to the number of keyword) (Raghavan & Wong, 1986). However, in spite of its obvious simplicity, the vector space model showed to be at least as good other models in terms of quality of results and calculation speed.

2. Texts pre-processing

Due to the morphological complexity of the Arabic language, this last much became an integral part of systems of Arab information research (IR). Studies of the Arab information research showed that the use of Arab roots as terms of indexing has improve the efficiency of research substantially by contribution to the use of words (Al-Kharashi & Evens, 1994; Abu-Salem et al., 1999; Hmeidi et al., 1997). For the need of the Arab information research, we opted for a thin lemmatization (Kadri, 2003) that makes the truncation of a restricted set of affixes. Our process of lemmatization does not lean on the syntactic dependence rules but on the morphological rules. We arrange two resources: a basis of affixes and a dictionary of lemmas. We test a word that belongs to the dictionary otherwise we operate a truncation of affix and we add the word and its lemma to the dictionary. In this paper we are not interested in UML Modeling of pre-processing stage, for more details refer to the most recent advance in UML framework for Arabic languages (Tahir *et al.*, 2004).

3. Hadith database indexing

Once the pre-processing is done on all texts of the basis, the indexing operation can begin. This stage consists in raising terms of texts in order to establish a link between every term (keys of the index) and texts that contain them. Once the terms are raised, a weight must be assigned to each of them for its presence in every text. The ponderation of terms used in *Authentique* is TFIDF. In (Fig. 1) a state Chart UML diagram is used for modeling the indexing process. This diagram permits modeling the behavior of class Index while insisting on the order of events and permits to visualize the different states of the text in the process of indexing since the obtaining of this text until the insertion of document information in the database.

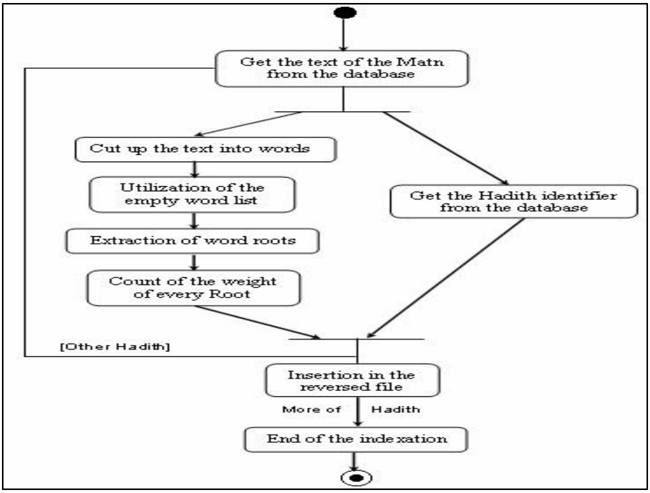

Get the text of the Matn from the database
Cut up the text into words
Utilization of the empty word list
Extraction of word roots
Count of the weight of every Root
Get the Hadith identifier from the database
[Other Hadith]
Insertion in the reversed file
More of Hadith
End of the indexation

Figure 1. *State chart Diagram for indexation*

4. Ponderation terms

Different formulations of the ponderation term are proposed in the literature (see for example (Salton & Buckley, 1988; Kwok, 1996). Only the one kept for *Authentique*, by reason of the quality of which it made proof in many systems, is presented in this section. The method used is the so-called TFIDF -Term Frequency, Inverse Document Frequency -. This formulation makes the hypothesis that a term is important for a given text, if it appears often in this text and that few texts contain it (Salton & Buckley, 1988). This ponderation is defined by (Eq. 1).

$$\begin{aligned} TFIDF(w,d) &= TF_{w,d} \bullet IDF_{w,d} \\ &= TF_{w,d} \bullet \left(\left(\log_2 \frac{ND}{F_{w,d}} \right) + 1 \right) \quad ; \qquad (1) \end{aligned}$$

with: *w* a term, *d* a document, $TF_{w,d}$ the number of apparitions of *w* in *d*, $DF_{w,d}$ the number of texts of the basis that contains *w* and *N* the total number of texts in the documents data base.

5. Modeling the research process

In this section, we represent the flux of information between the functional components identified in the search system. The UML Sequence Diagram is used to illustrate the interaction between the functional components and the flux of information through the system. The general architecture shows the connectivity between all functional components of the system, including processes of post-processing and research. Fig. 2 shows the global architecture.

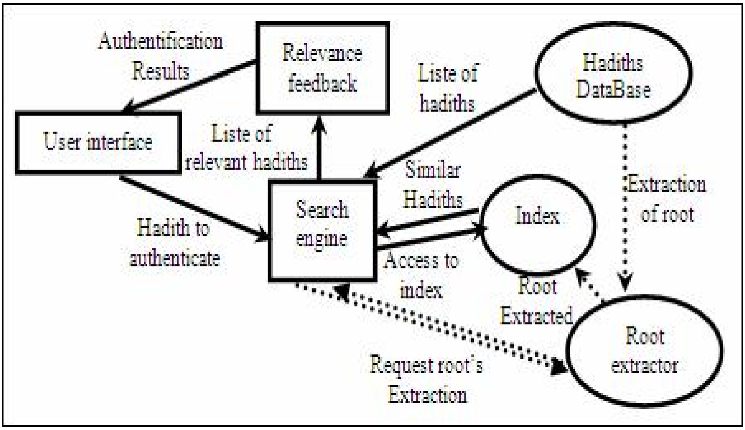


Figure 2. The general architecture of the research system

These components include parts of the system implied in the entrance of the user's queries, the processing of these queries, their passage through the system to generate a set of results, and finally the presentation of these results to the user.

5.1. The user interface

This component provides a multi-modal interfacing for the research of information permitting an interaction through the different modules of the software. This Interface is used for the formulation of queries, the presentation of results, and the enrichment of these queries in the light of these results.

5.2. The index

This component, include the reversed file that is used to store the result of indexation processes. It is implied in the research of the best similar documents of the initial query (this process will be described below).

5.3. The search engine

This component provides the kernel of functionalities of the research information system. It has access to the index to get the similar documents to the user's query. This functionality will be based on a vector space research model.

5.4. Relevance Feedback

This component is responsible for the result set classification (identification of documents) received from the component research engine on the basis of their relevance to the initial query. Documents judged relevant will be used to reformulate a new query while modifying weights of terms that appear in these documents. The relevant document list will be enriched therefore and the set of results will be sent to the Interface component to present them to the user.

5.5. UML sequence diagram

Fig. 3 presents the UML Sequence Diagram. This diagram show the different inter-connections between components of systems that have not been mentioned in the Fig. 2 (for example between the component Relevance feedback and Index to facilitate the enrichment of the query, and between the component Research engine and the component User Interface to facilitate the research of a hadith from the knowledge of its term.

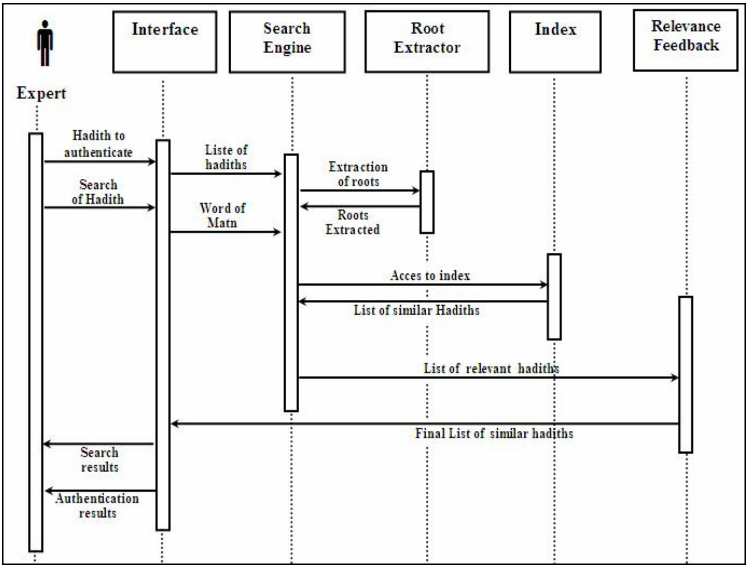


Figure 3. Diagram of sequence of the information flux in the system of hadiths research

6. Query treatment

The processing and the formulation of queries have been extensively studied. These studies concern the enrichment of queries for example by the automatic addition of terms (Attar & Fraenkel, 1977) thanks to the utilization of a thesaurus (Qiu & Frei, 1993). Terms of queries are pondered in the same way as terms of a document (the indexed texts).

7. Count of similarities

The index permits to get the list of documents that contains terms of the query. Similarities are calculated according to every term of the query by the hold in amount of the term weight in hadiths and in the query.

Cosine is one of the most frequently-used similarity measures (Salton, 1983). It consists in calculating values of cosines of angles separating vectors of hadiths and the vector of the query (Fig. 4). According to the vector space model, hadiths and the queries are represented in the same space. With regard to a simple scalar product, this measure presents the advantage to normalize scores of every hadith according to its size. The cosine measure is defined as follows (Eq. 2) :

$$\cos ine(d,r)=\frac{\sum_{w\in d\cap r} TFIDF_{w,d}\bullet TFIDF_{w,r}}{\sqrt{\left(\sum_{w\in d} TFIDF_{w,d}{}^{2}\right)\bullet\left(\sum_{w\in r} TFIDF_{w,r}{}^{2}\right)}}; \qquad (2)$$

With: *w* a term, *d* a document of the basis, *r* the query, $TFIDF_{w,d}$ the weight of *w* in *d* and $TFIDF_{w,r}$ the one of *w* in the query.

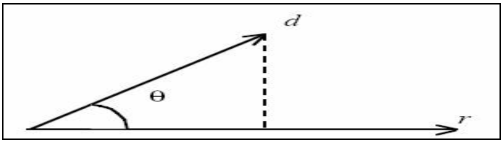


Figure 4. The cosine as measure of similarity process

8. Search Algorithm using cosine

The algorithm (Fig. 5) is used to calculate scores of documents of the database according to a query after labeling and lemmatization of this last.

- **For every term T of the query R:**
  - **Get the list of hadiths containing T, using the reversed file**
  - **For every hadith H of this list:**
    - **Update the score of H according to weights of T in the hadith and in the query:**

      Score (H) = score (H) + (weight (T, H)* weight (T, R))
    - **Update the sum of squares of weights used for the count of the score for H and for R (this is are used for the normalization of scores, they represent the norm of vectors of hadiths and the query),**
- **Normalize scores of every hadith;**
- **Order hadiths according to the normalized scores.**

Figure 5. Count of document scores according to Cosine according to a given query

9. Relevance feedback

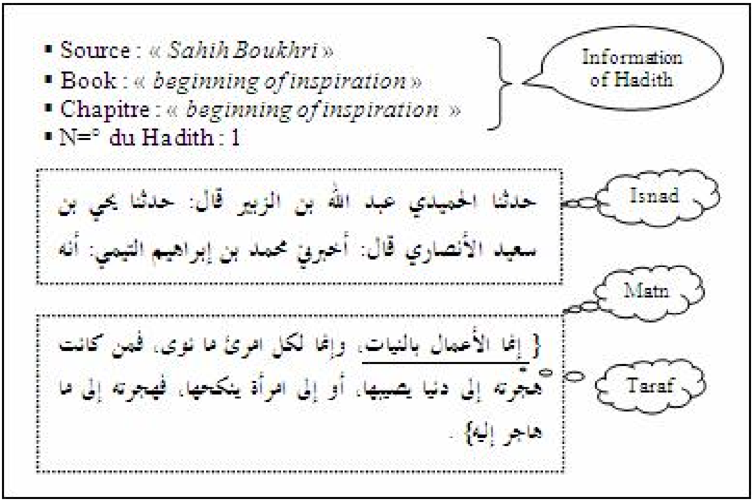
It is reasonable to think that a term that is frequently present in documents judged relevant and very little present in the documents judged not relevant is judged like a good term for a query. This remark can permit to increase the certain term weight in certain queries. If we considers that term t_i appears n_i time in the N documents of the collection and that it appears r_i time in the R (by default R=5) relevant documents of this same collection, the count of weight w_i is defined by (Eq. 3) :

$$w_i \approx \log \frac{r_i \bullet (N - n_i - R + r_i)}{(R - r_i) \bullet (n_i - r_i)} \qquad ; \qquad (3)$$

All terms found in *R* documents are classified in the decreasing order by weight of relevance *w*$_i$. Weights of the *K* (K by default = 10) first terms are calculated and they are merged then with terms of the initial query to create a new query. Some terms among the selected terms can be in the initial query. For the first selected terms that are not in the initial query, the weight is put to 0.5. For those that are in the initial query, the weight is put to 0.5**TFt*, where *TFt* is the frequency of term t_i apparition in the initial query. The selected terms are merged with the initial query to formulate an enriched query .

10. Experimental Results

A *hadith* or word of the Prophet (Peace and Blessings be upon Him) is composed of three parts (Fig. 6): *Matn*, *Isnad* and *Taraf*. For the need of the authentication we are only interested by the *Matn* part .



- Source : « *Sahih Boukhri* »
- Book : « *beginning of inspiration* »
- Chapitre : « *beginning of inspiration* »
- N=° du Hadith : 1

حدثنا الحميدي عبد الله بن الزبير قال: حدثنا يحي بن
سعيد الأنصاري قال: أخبرني محمد بن إبراهيم التيمي: أنه

{ إنما الأعمال بالنيات، وإنما لكل امرئ ما نوى، فمن كانت
هجرته إلى دنيا يصيبها، أو إلى امرأة ينكحها، فهجرته إلى ما
هاجر إليه}.

Figure 6. Components of the Hadith

Takhridj El hadith or *hadith* authentication is an operation achieved by experts of the domain; which consists of constructing the tree of hadiths paths. This tree (Fig. 7) compares the different hadiths and allows the study of their respective degrees of veracity .

The textual data mining is exploited in the semantic resemblance discovery context between the *Matns* texts of the different hadiths. The utilization of this technique is going to help the expert in his task of authentication by an automatic research tool .

In order to evaluate our results on corpora in Arabic language, we chose a database that contains texts of about of 60 "*Hadiths*" (Harrag, 2005).

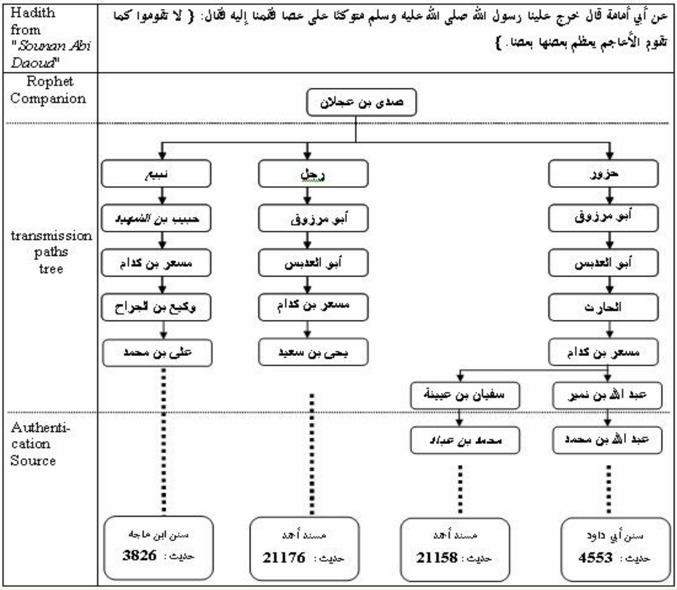


Figure 7. Result of the process of authenticity

10.1. The query

An example of query is represented by the need of authentication of the hadith n° 3974 of the source *"Sahih Musslem"* that concerns names, forenames and appellations (الكـــنى) of the Prophet (Peace and Blessings be upon Him) . The two roots extracted from the text of the *Matn* are: "سما" for "تسموا , باسمي " and "كنا" for "بكنـــيتي , تكنـــوا", these two roots appear in the query two times for each of them. So, the count of the weight according to the *TFIDF* weighting is made on a set of 54 hadiths, root "سما" appears in 19 hadiths whereas root "كنـــا" appears in 6 hadiths. Root "كنـا" is a weight of 7.41 (Eq. 4) although it only appears 12 times in the 6 hadiths with regard to root "سما" that appears 29 times in the 19 hadiths (Eq. 5), this is explained by the fact that the *TFIDF* measure gives a big discriminatory power to the very rare terms in the basis of hadiths .

The text (*Matn*) of the hadith n° 3974 of the *"Sahih Musslem"* is :

{ تسموا باسمي ولا تكنوا بكنيتي } .

$$\begin{aligned}\mathrm{TFIDF}(\text{"kana"},3974) &= \mathrm{TF}_{\mathrm{kana},3974} \bullet \mathrm{I}DF_{kana,3974}\\ &= 2\bullet\left(\left(\log_2\frac{54}{06}\right)+1\right) \qquad ; \qquad (4)\\ &= 7.41\end{aligned}$$

$$\begin{aligned}\mathrm{TFIDF}(\text{"sama"},3974) &= \mathrm{TF}_{\mathrm{sama},3974} \bullet \mathrm{I}DF_{sama,3974}\\ &= 2\bullet\left(\left(\log_2\frac{54}{19}\right)+1\right) \qquad ; \qquad (5)\\ &= 3.71\end{aligned}$$

The results of weighting are represented in Table 1. :

Term of query	Weight TFIDF
كنا	7.41
سما	3.71

Table 1. Example of query

10.2. Count of similarities

Terms index of relevant hadiths for the previous query are given with their degrees in the Table 2. ; Table 3.

represents the score of the list of hadiths given by the *Authentique* system in the purpose of authentication operation. The text (*Matn*) of the hadith n° 3974 of the *"Sahih Musslem"* is considered like a query for the process of research : Hadith N=° 3974:

{ تسموا باسمي ولا تكنوا بكنيتي } .

The given hadiths are: hadiths N=° 3976, 3977, 3978, 3979 of the *"Sahih Musslem"*:

Hadith N=° 3976:

{ تسموا باسمي ولا تكتنوا بكنيتي فإنما أنا قاسم أقسم بينكم }

Hadith N=° 3977:

{ سموا باسمي ولا تكنوا بكنيتي فإنما بعثت قاسما أقسم بينكم }

Hadith N=° 3978:

{ تسموا باسمي ولا تكنوا بكنيتي فإنما أنا أبو القاسم أقسم بينكم }

Hadith N=° 3979: { سموا باسمي ولا تكتنوا بكنيتي }

Term of index	Weight for the hadith 3976	Weight for the hadith 3977	Weight for the hadith 3978	Weight for the hadith 3979
كنا	7.41	7.41	7.41	7.41
سما	12.78	3.71	3.71	3.71
بعث	0.00	17.25	0.00	0.00
قسم	10.12	10.12	5.06	0.00

Table 2. Index Terms of hadiths relevant for R = 3974

Term of index	Score of hadith 3976	Score of hadith 3977	Score of hadith 3978	Score of hadith 3979
كنا	54.90	54.90	54.90	54.90
سما	48.15	13.76	13.76	13.76
Cos (d, r)	0.68	0.38	0.85	1.00

Table 3. Score of hadiths

Count of the Score of the hadith 3976 of the *"Sahih Musslem"* (Eq. 6):

$$\text{Score} = \frac{((7.41*7.41)+(12.98*3.71))}{\sqrt{\left((7.41)^2+(3.71)^2\right)\bullet\left((7.41)^2+(12.98)^2\right)}} \; ; \qquad (6)$$

$= 0.68$

10.3. Relevance feedback

For the query of the example of the previous section, the set of relevant hadiths is 4 on 54 of the hadiths basis. The count of the relevance weight is given in the table 5.

Term of index	Weight of relevance	Term of query	Weight of selection	Term of query	Weight of expansion
قسم	2.16	قسم	0.5	قسم	0.5
كنا	1.36	كنا	1.0	كنا	8.41
سما	0.44	سما	1.0	سما	4.71

Table 4. Example of enrichment of the query 3974

Count of the relevance weight for term "قسم " (Eq. 7):

$$w \approx \log \frac{7 \bullet (54 - 7 - 4 + 6)}{(4 - 6) \bullet (7 \;\; 6)} \qquad ; \qquad (7)$$

$$\approx 2.16$$

We put K=3, the selected terms are: "قـــسم", "كنـــا", "سمـــا", their weights of selection are given in the table 5. This table contains weights of terms of the query enriched; the term added to the initial query is the term: "قـــسم", its weight is put to 0.5 and weights of the two other terms that exist in the initial query are modified .

After the application of the same count of the score for a second time the system is going to enrich the list of results by a set of new *hadiths*. This set can be divided in two class of hadiths: relevant and not relevant .

The new *hadiths* list given by the enrichment of the query is

Relevant hadiths list :

Hadith N=° 3975 **:**

{ إن أحب أسمائكم إلى الله عبد الله وعبد الرحمن }

Hadith N=° 3984 **:**

{ لا تسم غلامك رباحا ولا يسارا ولا أفلح ولا نافعا }

Not relevant hadiths list :

Hadith N=° 3975 **:**

{ أقسموا المال بين أهل الكتاب على كتاب الله }

11. Research system evaluation

Since the objective of a research system is to minimize the effort provided by a user to reach the goal, then criteria of evaluation are interested by the search speed and the quality of results. Among objectives of a research system, there are the reductions of the number of documents not relevant found by error and the reduction of the number of relevant documents not reported. Criteria corresponding to these two objectives are named precision and recall .

Let's take the query of the example of the previous section, the relevant hadiths set given is 8 hadiths on 10 relevant hadiths, and the number of all hadiths given as answer of the query is 12 hadiths. Therefore values of the two measures of evaluation precision (Eq. 8) and recall (Eq. 9) are :

$$\text{Precision} = \frac{\text{number of relevant hadiths found}}{\text{number of hadiths found}} = \frac{8}{12} = 0.66 \quad ; \qquad (8)$$

The value of noise is therefore: $Noise = 0.34$.

$$\text{Recall} = \frac{\text{number of relevant hadiths found}}{\text{number of relevant hadiths to find}} \quad ; \quad \frac{8}{10} = 0.80 \qquad (9)$$

The value of Silence is therefore: $Silence = 0.20$

Conclusion

In this article, we used the UML as a modeling tool for an automatic text mining system, called A*uthentique*, for knowledge extraction from a databases of Prophetic Traditions "*Hadiths*". This system provides a list of classified *hadiths* according to their degrees of similarity with respect to user's query. The implemented methods of text mining in *Authentique* are classical methods such as vector space model, TFIDF, and cosine measure. These have been chosen in order to assure the quality of hadiths lists considered a set of Arabic texts to be ordered by decreasing order of relevance, *i.e.* from the most to the less relevant. The use of these techniques yielded efficiency on large textual databases in Arabic language, thus extending works that have been done so far in Western Languages such as French and English. The results obtained confirm the independence of the statistical methods from the language.

Futur work consists in the study of the new methods of classification and segmentation of knowledge in the textual data bases. These methods represent a better structuring process.

References

Abu-Salem, Hani, Al-Omari, M. and Martha, E. (1999) "Stemming Methodologies Over Individual Query Words for Arabic Information Retrieval". JASIS, Vol 50 No 6, pp. 524-529.

Al-Kharashi, I. and Martha, E. (1994) "Comparing Words, Stems, Roots and ace Index Terms in year Arabic Information Retrieval". JASIS, Vol 45 No 8, pp. 548-560.

Attar, R. and Fraenkel, A.S. (1977) "Local feedback in full-text retrieval systems". Journal of the ACM, Vol 24 No 3, pp. 397-417.

Bilhaut, M. F. (2006), "Analyse automatique des structures thématiques discursives: Application à la recherche d'information", Phd Thesis, University of CEAN/ Basse-Normandie, French, pp. 19-20

Harrag, F. (2005) "Modélisation d'expertise dans les base de données: application à l'authentification des Traditions Prophétiques Hadith". Mémoire de Magister, (M.Sc. Dissertation), Université Farhat Abbas, Sétif, Algérie.

Harrag, F. and Hamdi-Cherif, A. (2006) "Modélisation UML de la fouille des textes en langue arabe et application à l'authentification des Traditions Prophétiques Hadith". Conférence Internationale sur l'Informatique et ses Applications CIIA06, Saida, Algérie, Vol 1 No 1, pp. 27.

Hmeidi, Ismail, Ghassan, K. and Martha, E. (1997) "Design Implementation and of Automatic Indexing for Information Retrieval with Arabic Documents". JASIS, Vol 48 No 10, pp. 867-881.

Kadri, Y. (2003) "Research of information translinguistique on documents in Arabic". Oral prédoc Report, DIRO, University of Montreal.

Kwok, K.L. (1996) "new method of weighting query terms for ad-hoc retrieval". Acts of ACM/SIGIR'96 Conference one Research Development and in Information Retrieval, Zurich, Switzerland, pp. 187-195.

Qiu, Y. and Frei, H.P. (1993) "Concept based query expansion". Acts of ACM/SIGIR'93 Conference one Research Development and in Information Retrieval, Pittsburgh PA, USA, pp. 160-169.

Raghavan, V.V. and Wong, S.K.M. (1986) "a critical analysis of vector space model for information retrieval". Newspaper of the American Society for Information Science, Vol 37 No 5, pp. 279-287.

Salton, G. (1983) "Introduction to Modern Information Retrieval". Mc Graw-Hill, New York, USA.

Salton, G. and Buckley, C. (1988) "Term weighting approaches in automatic text retrieval", Information Processing and Management, Vol 24 No 5, pp. 513-523.

Tahir, Y., Chenfour, N. and Harti, M. (2004) "Modélisation à objets d'une base de données morphologique pour la langue arabe". JEP-TALN04 Traitement Automatique de l'Arabe, Fès, Morroco.