

UML modeling of text mining in Arabic language Application to the Prophetic Traditions "*Hadiths*"

Fouzi Harrag

*Computer science department, University Ferhat Abbas
19000 - Sétif, Algeria
Hfouzi2001@yahoo.fr*

&

Aboubekour Hamdi-Cherif

*Computer College
Qassim University
PO Box 103 – 51411 Buraydah
Saudi Arabia
elhamdi62@ksu.edu.sa*

Introduction

- Integration, modeling, structuring and extracting knowledge are crucial for different domains :
 - medical, biological.
 - linguistic, sociological.
- **Why study *Hadith* ?**
 - **What are Prophetic Traditions *Hadith* ?**
 - Word of the Prophet (Peace and Blessings be upon Him).
 - The way to prophetic knowledge
 - *Sunna* is the second source of Islamic Legislation
- **Goal**
 - To provide an easily-searchable list of Prophetic Traditions *Hadith*
 - Ordered according to degree of similarity
- **Purpose**
 - Propose a system of text mining for the automatic authentication of Hadith
 - Degree of veracity ().
- **Tool : Vector space model.**

Overview

1. Text mining and vector space model.
2. UML Modeling of *Authentic* System.
3. Authentication of Prophetic Traditions.
4. Experimental Results.
5. Conclusion.

Text Mining

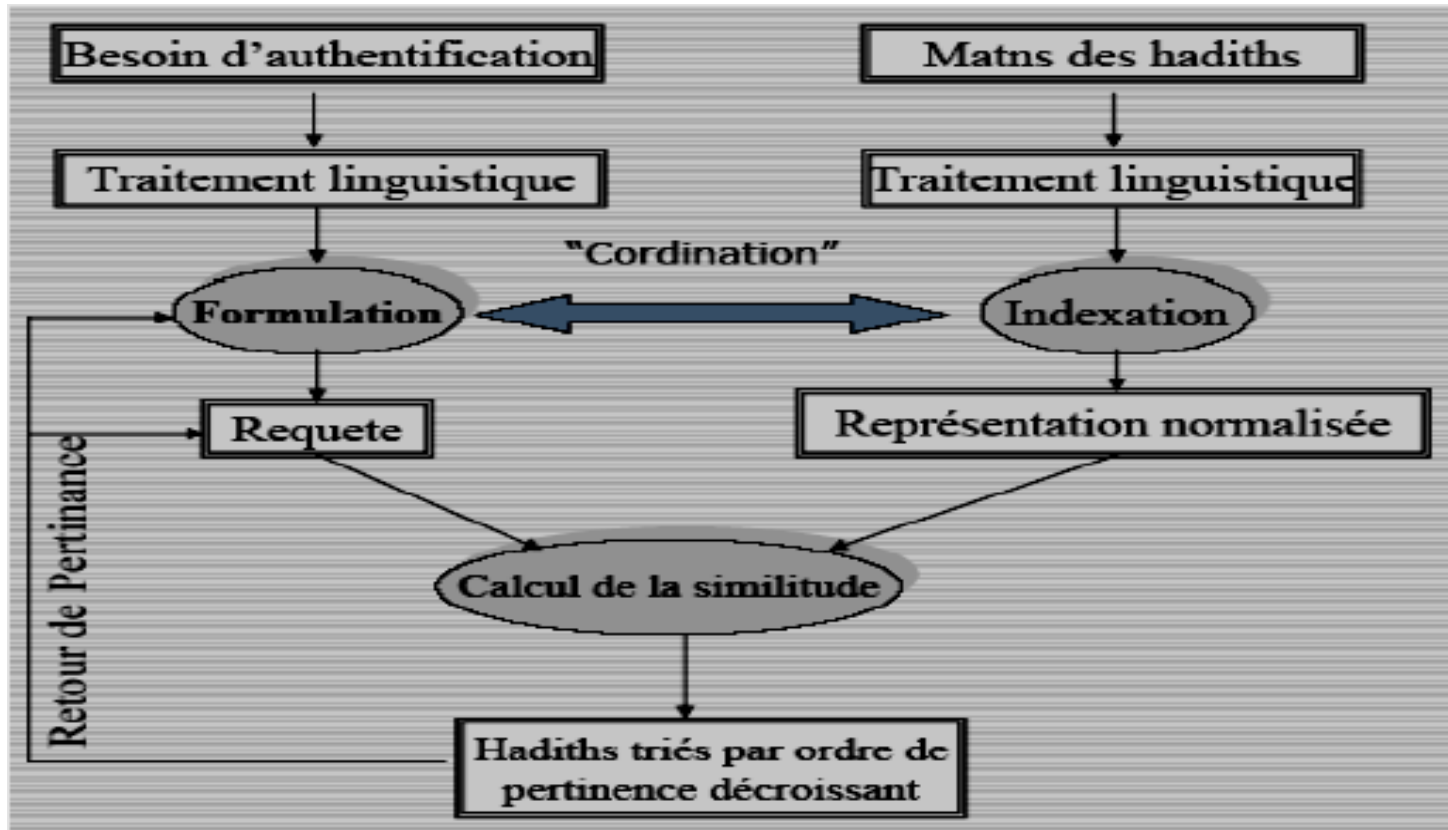
- The process of text mining can be represented according to :
 - **A vector space model.**
 - **A probabilistic model.**
 - **A hybrid model.**
- Its realization can be made of :
 - **The own statistical methods of the vector space model;**
 - **The fuzzy logic or the modal logic;**
 - **The hidden Markov model;**
 - **others.....**

Text mining algorithm

- **Database Indexing**
 1. Recognition of terms.
 2. Empty word elimination.
 3. Lemmatisation or Stemming.
 4. Assignment of weights (Reversed File).
- **Research of documents relevant to request**
 - Idem pour de 1 à 4 (without Reversed File).
 - Count of similarities.
- **Presentation of research results.**

The vector space model

- The vector space model chosen as basis in *Authentique* research system .



Indexing and research of documents in the vector space model

Post processing

- Prophetic Traditions = Texts written in Arabic language.
- Morphological treatment of the Arabic language.
- Use of Arabic root as indexing terms.
- Process of lemmatisation :
 - By dictionary.
 - By truncation.

Indexing: Process of indexing

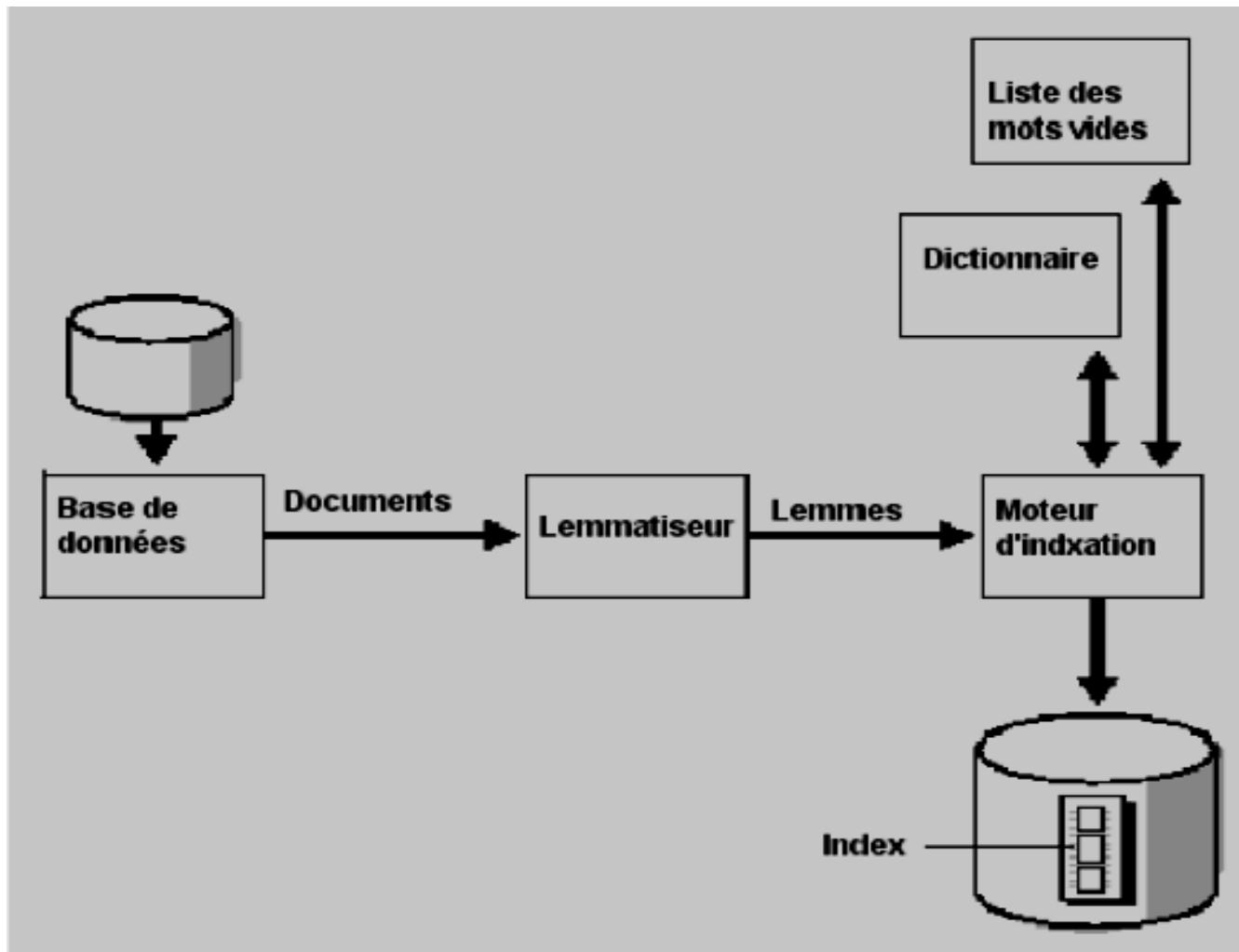
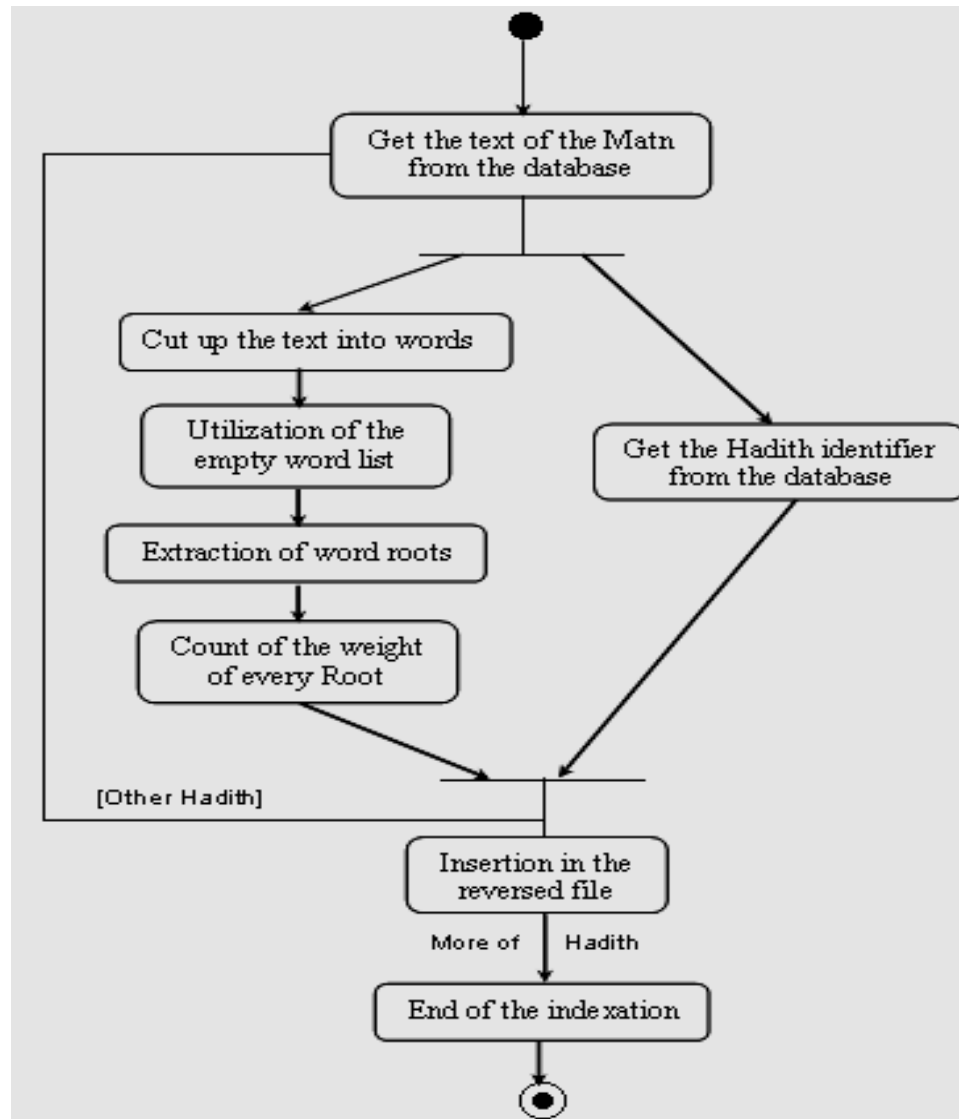


Diagram of the indexing process

Indexing : Modeling



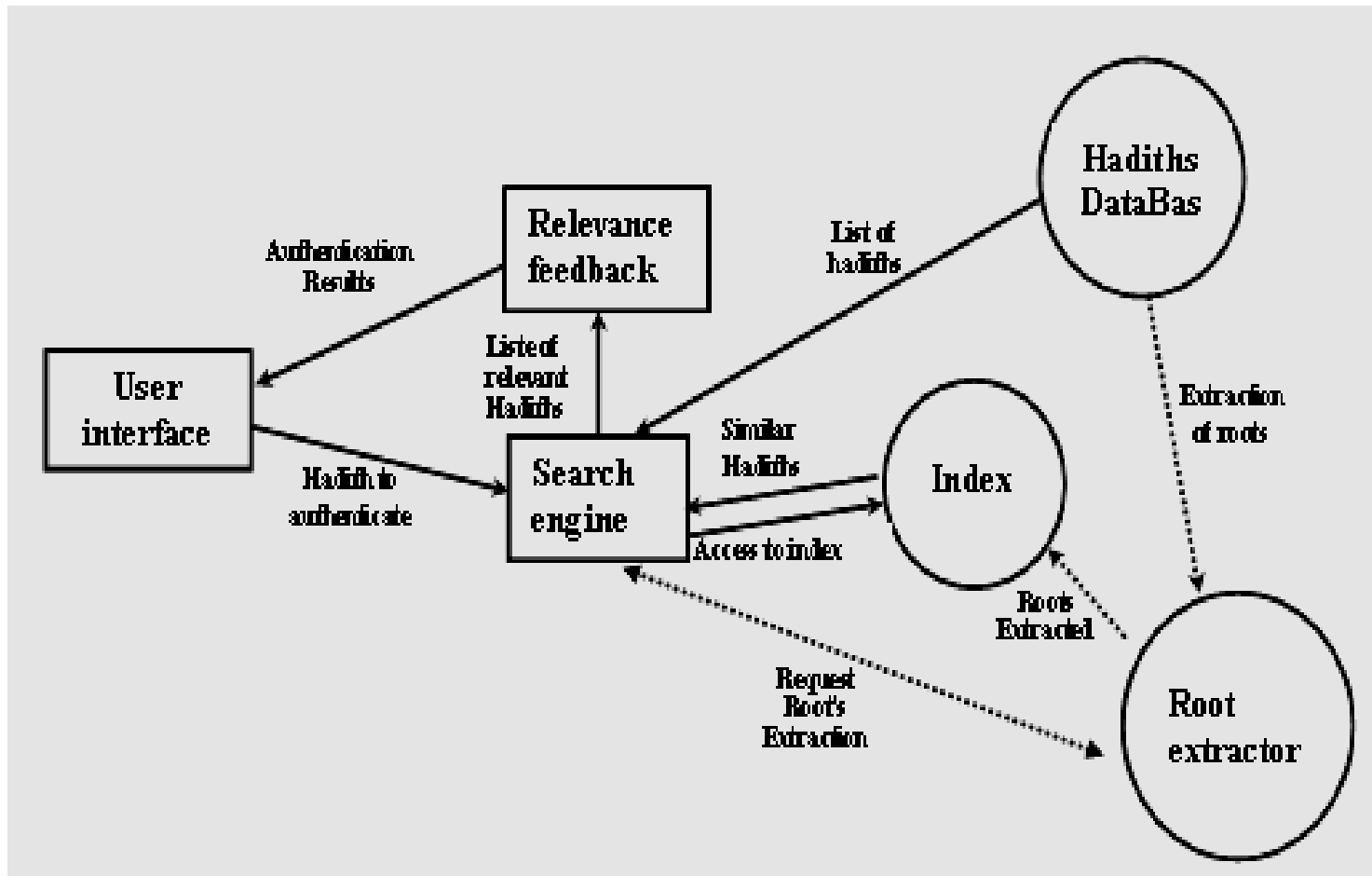
Indexing: Ponderation of terms

TFIDF : Term Frequency, Inverse Document Frequency

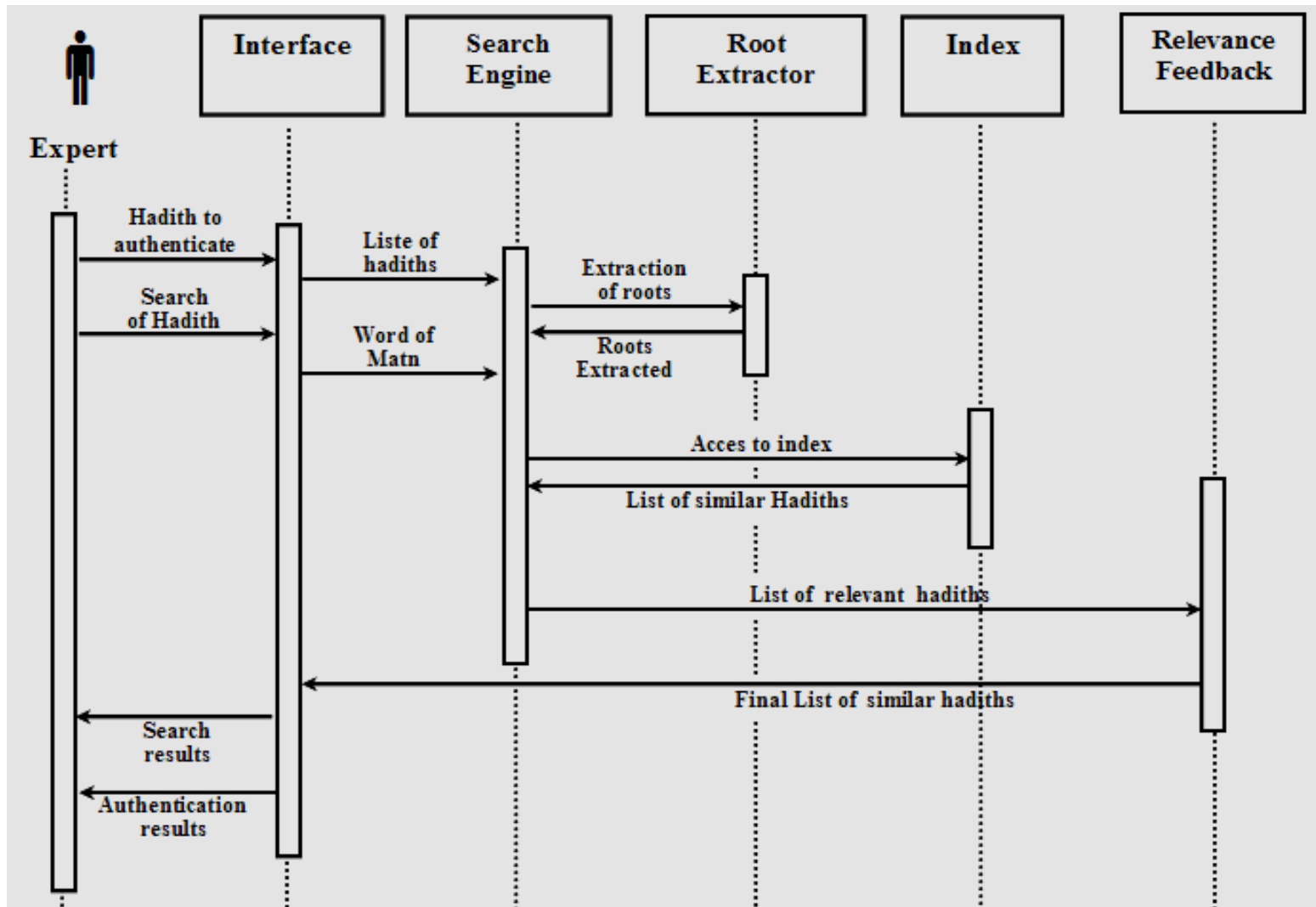
$$\begin{aligned} \text{TFIDF}(w, d) &= \text{TF}_{w,d} \cdot \text{IDF}_{w,d} \\ &= \text{TF}_{w,d} \cdot \left(\log_2 \frac{N}{\text{DF}_w} + 1 \right) \end{aligned}$$

The TFIDF measure gives a discriminatory power to the very rare terms in the textual data base.

The general architecture of the Research system

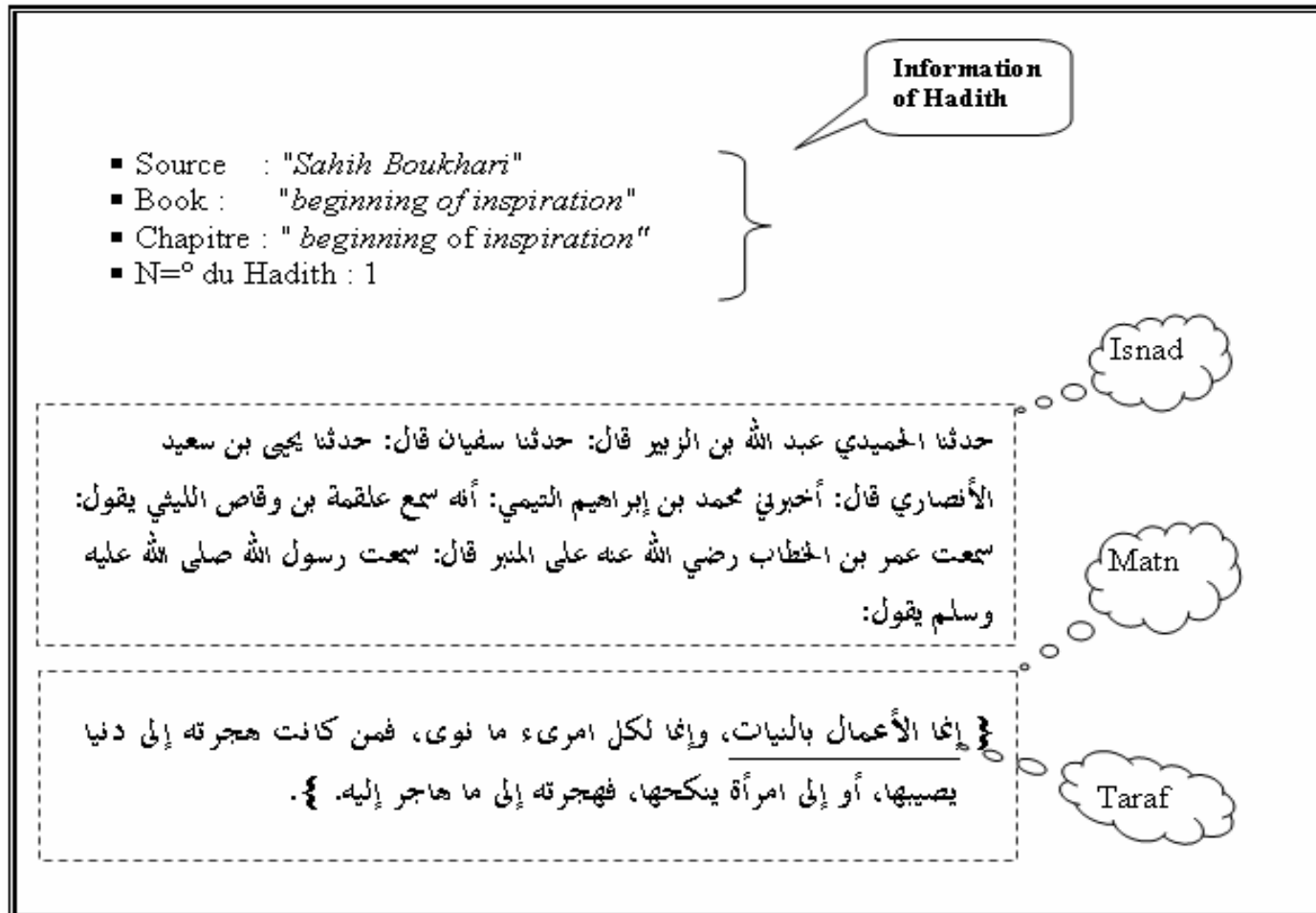


Information flux sequence diagram



Hadith Authentication

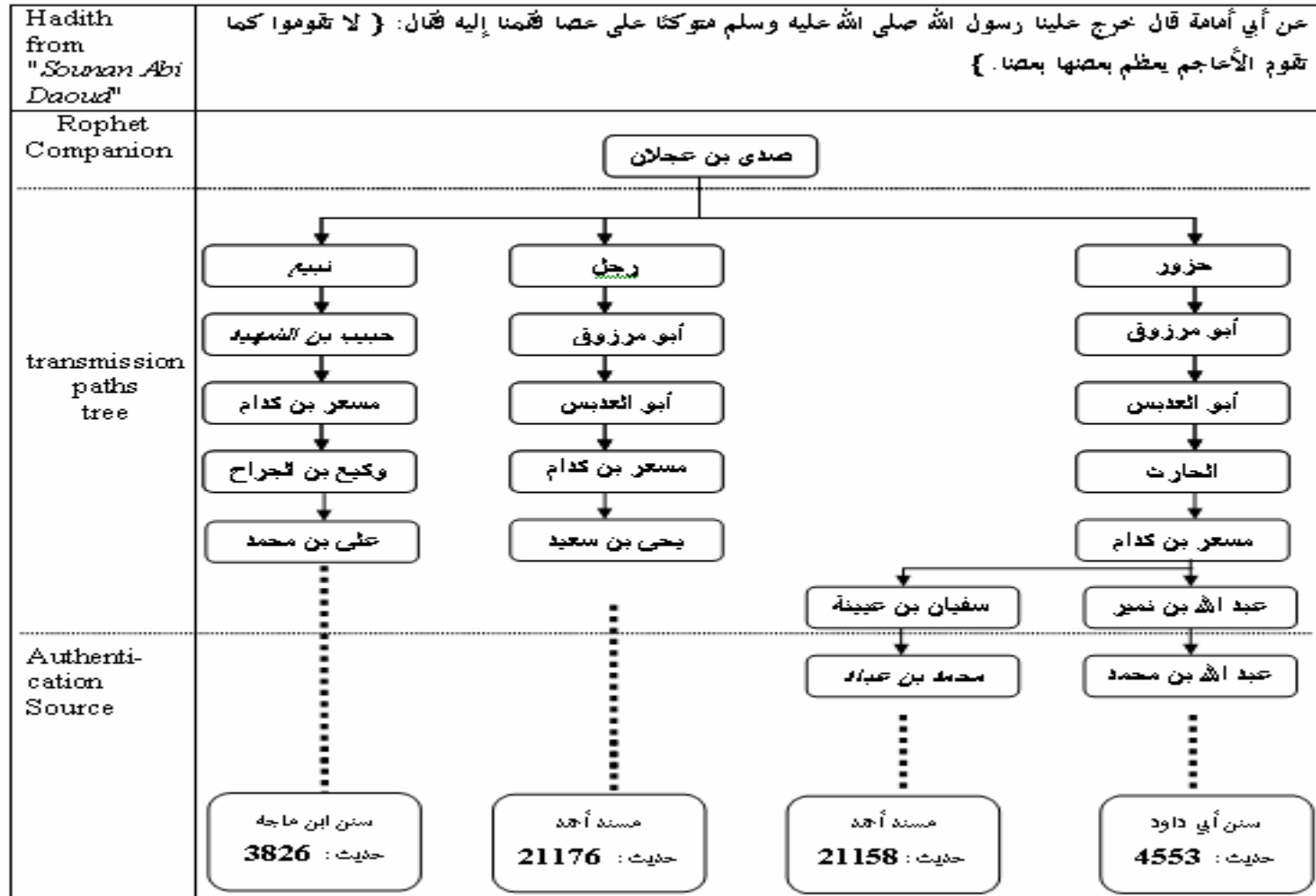
- *Hadith* = 3 parts : *Matn*, *Isnad* et *Taraf*.



Why authentication of "Hadith" ?

- Constructing the tree of the chain of transmission or narration of *Hadith*.
- The *Hadith* expert is interested in :
 - accessing various compendia of *Hadith*,
 - finding a given *Hadith*, or similar,
 - comparing different chains of transmission,
 - knowing the veracity degree.

Authentication results



Text mining applied to authenticate *Hadith*

- Text mining = text content analysis
- Used in discovering semantic similarity.
- Allows automatic search
- Helps expert in finding texts and in establishing veracity degree.

Queries : terms ponderation

Simple evaluation

- Arabic Corpus : Data base of 60 *Hadith*

- Query n° 3974 of "*Sahih Musslem*":



“

- $TFIDF("كنا", 3974) = TF_{,3974} * IDF_{,3974} = 2 * ((\log_2 54/06) + 1) = 7.41.$
- $TFIDF("كنا", 3974) = TF_{,3974} * IDF_{,3974} = 2 * ((\log_2 54/19) + 1) = 3.71.$

Terme de la requête	Poids (TFIDF)
كنا	7.41
كنا	3.71

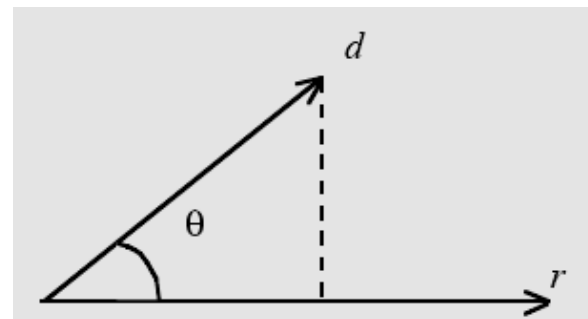
Similarities calculation

Research : calculates a measure of similarity between each hadith of the data base and query made by the expert.

Cosine :

calculates cosine values of angles between vectors of hadiths and the vector of query

$$\text{cosine}(d,r) = \frac{\sum_{w \in d \cap r} TFIDF_{w,d} \cdot TFIDF_{w,r}}{\sqrt{\left(\sum_{w \in d} TFIDF_{w,d}^2 \right) \cdot \left(\sum_{w \in r} TFIDF_{w,r}^2 \right)}}$$



Terme de l'index	Poids pour le hadith 3976	Poids pour le hadith 3977	Poids pour le hadith 3978	Poids pour le hadith 3979
كنا	7.41	7.41	7.41	7.41
سما	12.98	3.71	3.71	3.71
بعث	0.00	17.25	0.00	0.00
قسم	10.12	10.12	5.06	0.00

List of relevant Hadiths

- Hadith N=° 3976 : ❁ “
- Hadith N=° 3977 : ❁ “
- Hadith N=° 3978 : ❁ “
- Hadith N=° 3979 : ❁ “

Terme de l'index	Score du hadith 3976	Score du hadith 3977	Score du hadith 3978	Score du 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

Relevance result

Relevance weight : •

$$w_i \simeq \log \frac{r_i(N - n_i - R + r_i)}{(R - r_i)(n_i - r_i)}$$

According to whether or not the term is part of the query •

Selection weight : ■

$$W_s = 0.5 * TFt$$

o ù

$$W_s = 0.5$$

Expansion weight : ■

$$W_e = W_t + W_s$$

o ù

$$W_e = W_s$$

Query Enrichment

Let: $K=3$ selected terms are : " ", " ", " ",
 added term : " " : weight =0.5.

Terme de l'index	Poids de pertinence
قسم	2.16
كنا	1.36
سما	0.44

Terme de la requête	Poids de sélection
قسم	0.5
كنا	1.0
سما	1.0

Terme de la requête	Poids d'expansion
قسم	0.5
كنا	8.41
سما	4.71

List of relevant hadith :

Hadith N=° 3975 : ☼

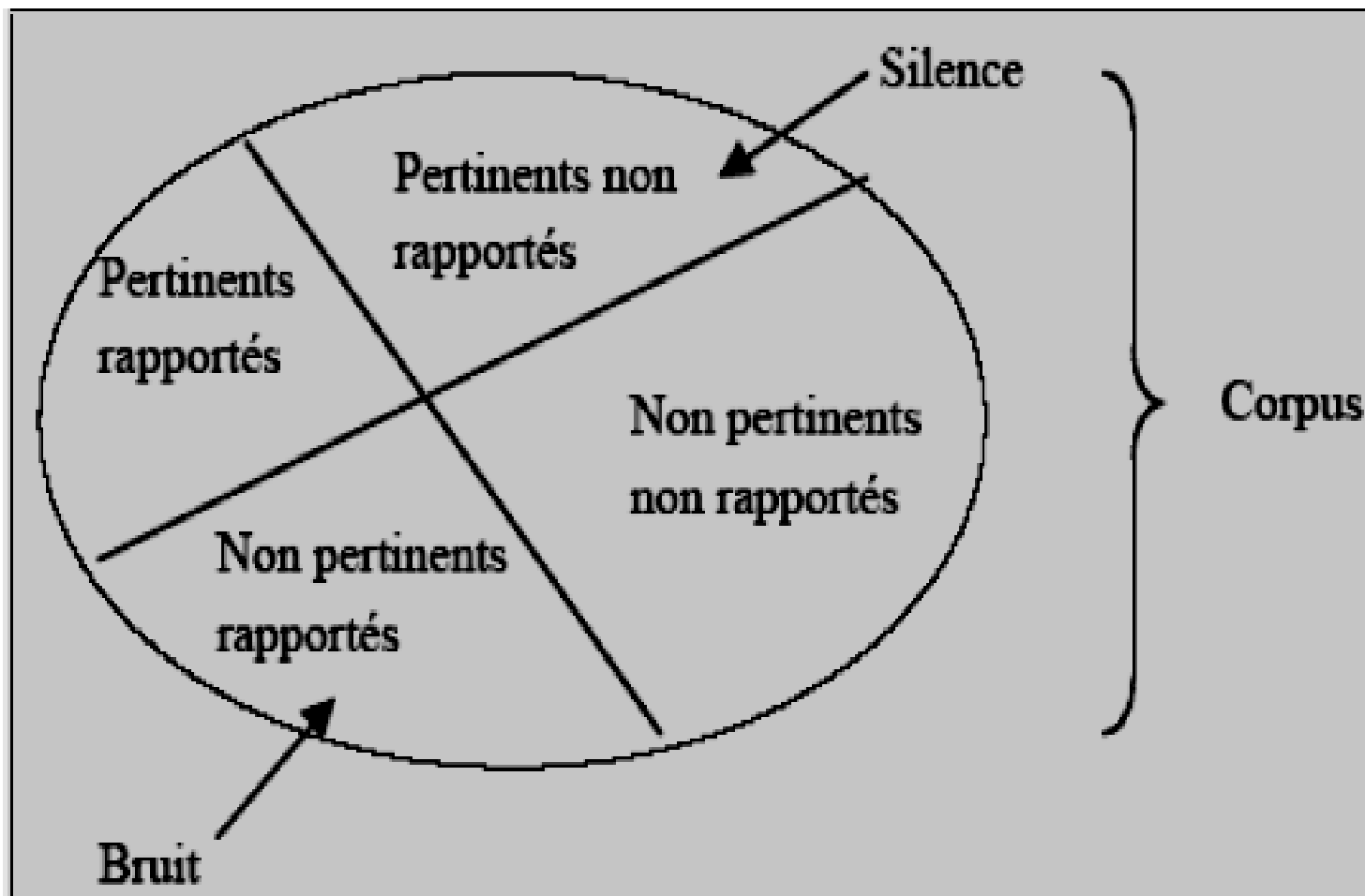
Hadith N=° 3984 : ☼

List of irrelevant hadith :

Hadith N=° 3997 : ☼



Research system evaluation



Research system evaluation

$$\text{Precision} = \frac{\text{number of relevant hadiths found}}{\text{number of hadiths found}}$$
$$= \frac{8}{12} = 0.66$$

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

$$\text{Recall} = \frac{\text{number of relevant hadiths found}}{\text{number of relevant hadiths to find}}$$
$$\frac{8}{10} = 0.80$$

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

Indexing in Authentique

Interface de Authentification Automatique

Recherche Indexation Liste des Documents

Indexer Supprimer Quitter

Unités Lexicales: 86

Unité Cherchée: [] Atteindre

Nombre de Documents: 10

Fréquence Globale: 10

Numéro	Nbre Occurrence	Poids
10	1	3.04
9	1	3.04
8	1	3.04
7	1	3.04
6	1	3.04
5	1	3.04
4	1	3.04
3	1	3.04
2	1	3.04
1	1	3.04

Indexer la base

Type de construction

- Globale (reconstruire toute la base)
- Sélectif (seulement les hadiths sélectionnés)

Liste des Hadiths

Choisir la liste des Hadiths à indexer :

Debut : 1

Fin : 921

Information du Hadith :

Hadith N° : 1

Texte du Hadith :

إِذَا الْأَعْمَالُ بِالنِّيَّاتِ وَإِنَّمَا لِكُلِّ امْرِئٍ مَا نَوَى فَمَنْ
كَانَتْ هِجْرَتُهُ إِلَى دُنْيَا يُصِيبُهَا أَوْ إِلَى امْرَأَةٍ يَنْكِحُهَا
فَهِجْرَتُهُ إِلَى مَا هَاجَرَ إِلَيْهِ

Traitement du Hadith :

- Segmentation
- Elimination des Mots Vides
- Lemmatisation
- Assignment des Poids

segmentation terminée

0%

0%

Créer Stop Quitter

Research in Authentique

The screenshot displays the 'Interface de Authentification Automatique' window. It features three tabs: 'Recherche', 'Indexation', and 'Liste des Documents'. The 'Recherche' tab is active, showing search parameters and results.

Information de la Requête :

- Requete N° : 4553
- Source : سنن أبي داود
- Livre : كتاب الأدب
- Chapitre : باب في قيام الرجل للرجل

Texte de la Requête :

خرج علينا رسول الله صلى الله عليه وسلم
متوكئا على عصا فقمنا إليه فقال لا تقوموا كما
تقوم الأعاجم يعظم بعضها بعضا

Traitement de la Requête :

- Segmentation
- Elimination des Mots Vides
- Lemmatisation
- Assignation des Poids

0%

Démarrer

Morphologie et Lemmatisation :

- Mots Vides**
- Rechercher**
- Dictionnaire**

Texte Lemmatisé de la Requete :

(Empty text area)

Index de la Requete (Nombre Occurrence ; Poids) :

Terme	Nbre Occurrence	Poids
الله	2	4.07
رسول	1	2.55
سلم	1	2.9
قول	1	4.04
خرج	1	4.36
وكئا	1	6.36
عجم	1	4.77
عظم	1	4.36
بعض	2	9.55

Quitter

The Windows taskbar at the bottom shows the 'démarrer' button, several application icons (including 'RB'), and the taskbar title 'Microsoft PowerPoint ... AUTHENTIQUE'. The system tray on the right shows 'FR', a volume icon, a network icon, and the time '8:13'.

Research Results

Interface de Authentification Automatique

Recherche

Information de

Requete N° :

Source :

Livre :

Chapitre :

Texte de la Re

Traitement de

Segmentat

Eliminatio

Lemmatisa

Assignatio

Morphologie et

Mots

Dict

Résultats de La Recherche

Contenu de la Réponse Sélectionnée :

Terme	Fréquence	Poids
الله	2	4.07
رسل	1	2.55
سلم	1	2.9
قول	1	4.04
خرج	1	4.36
وكأ	1	6.36

Documents Correspond au Terme :

Numéro	Tarkim	Score
750	21158	0.6203
780	21176	0.5940
650	3826	0.3589
320	614	0.1838
920	1580	0.1838
3	4682	0.1513
510	1383	0.1504
610	1785	0.1290
2	3530	0.1020

Nombre de Réponses : 19

Numéro	Tarkim	Score
750	21158	0.6203
780	21176	0.5940
650	3826	0.3589
320	614	0.1838
920	1580	0.1838
3	4682	0.1513
510	1383	0.1504
610	1785	0.1290
10	1882	0.1020
7	3383	0.1020
2	3530	0.1020

Contenu de la Réponse Sélectionnée :

Terme	Fréquence	Poids
الله	3	6.11
رسل	1	2.55
سلم	1	2.9
رحم	1	4.36
دعا	1	4.36
قبل	1	3.36
قول	4	16.14

خرج علينا رسول الله صلى الله عليه وسلم وهو متوكئ على عصا فقمنا اليه فقال لا تقوموا كما تقوم الأ عجم يعلم بعضها بعضا قال فكانا اشتبهنا أن يدعو الله لنا فقال اللهم اغفر لنا وارحمنا وارض عنا وتقبل منا وأدخلنا الجنة ونجنا من النار وأصلح لنا شأننا كله فكانا اشتبهنا أن يزيدنا فقال قد جمعت لكم الأمر

Évaluation des Résultats :

Nbre de Documents Pertinents Trouvés:

Nbre de Documents Pertinents à Trouver :

Pourcentage de Pertinence :

0 50 100

K Premiers Termes Pertinents :

Retour de Pertinence

Interface Utilisateur

Supprimer La Liste

Retour à La Recherche

Quitter

démarrer RB Microsoft PowerPoint ... AUTHENTIQUE FR 8:15

Authentication Results

AUTHENTIQUE

Affichage Recherche Authentification

Arbre des Chaînes de Transmissions

Choix d'une Source

- البخاري
- مسلم
- الترمذي
- النسائي
- أبو داود
- أحمد
- ابن ماجه
- الدارمي
- مالك

Source : سنن أبي داود Hadith N° : 4553 Référence : مرغوع Nature Chaîne : متصل Fermer

démarrer RB Microsoft PowerPoint ... AUTHENTIQUE FR 8:15

Conclusion

- UML modeling of authentication process of Prophetic Traditions or "***Hadith***".
- Design and Development of text mining system : ***AuthenTique***.
- Implementation of a method of knowledge extraction based on the vector space model.
- Confirmation of the statistical method independence opposite of the used language.
- Production of an original and unique synergy between ***Hadith***, UML and Textual Data mining.

Futur prospects

- Widening of the research system to make an universal text mining tool.
- Increased automation of the authentication of "*Hadith*" basing on:
 - classification → Topic categorisation.
 - Segmentation → new count of resemblance.
 - automatic construction model of the "*Hadith*" based on the meaning.