Journal of Computer Science 6 (6): 660-665, 2010
ISSN 1549-3636
© 2010 Science Publications

Heuristic Lemmatization for Arabic Texts Indexation and Classification

[1]Faten Khalfallah Hammouda and [2]Abdelsalam Abdelhamid Almarimi
[1]Department of Informatics, Faculty of Economy and Management, Sfax, Tunisia
[2]Department of Informatics, Higher Institute of Electronics BaniWalid, Libya

Abstract: Problem statement: This study proposed a system based on a heuristic lemmatization for Arabic text indexation and classification. This research is needed for a lot of NLP applications such as the research of information and automatic abstract. This system was not related to any linguistic rule. The proposed method was limited to five different domains: Sports, medicine, politics, economics and agriculture. The main idea is collecting different texts related to the chosen domains and studying them by extracting the pertinent terms. **Approach:** Every entered text had the formatting stage in which we can remove some words and letters that do not have any importance for the meaning. After that, the frequencies' average is calculated to classify the text and its related domain. **Results:** The main finality of the System of Indexation and Classification of Arabic Texts (SICAT) is to classify finally an unknown text in its suitable domain. So, it's to detect the text theme. To do this task, we applied a method by pertinent terms correspondence. It is about testing the correspondence of all pertinent terms of the text to classify with the keywords of every domain of the corpus. The domain, that constitutes the majority of terms having a correspondence with terms of the text, represents the theme that we look for to classify our unknown text. **Conclusion:** It holds two main parts: the indexation and the classification. The indexation stage is composed of three main parts: the pre-learning, the lemmatization and the frequencies' calculation. The classification stage is composed of two main components: the extraction of keywords and classification of new text. We have made many tests of verification to test the validation of the system. The system performance was evaluated on the different chosen domains, achieves 90% precision and 85% recall.

Key words: Natural language processing, indexation and classification

INTRODUCTION

Due to the evolution of the documentation, one of the stakes of research is to develop the tools permitting to manage the documentary mass automatically. That is why, a long time ago that libraries, documentary centers, producers of databases, companies and laboratories of researches used specialists to index their documents to have an idea about the restrained information in the organized basis of a document whole indexed first and on second part, to permit the ulterior research of these information. The more the computer domain develop, the more the automation of this process becomes absolutely a necessity.

The stage of indexing aims to determine the adherence domain of an unknown text. The major objective of the indexing is to find the most important concepts in documents and to create a representative description while using these concepts. The existing techniques are often limited to a very specialized domain and the analysis is very complex. Thus, in practice, one rather looks for representatives of concepts. These representatives can be different shapes (simple words or multi-terms). The choice of these last depends essentially on the easiness of treatment as well as the precision of sense representation.

The indexation of a document can be done manually by reading its title, abstract, the summary and the conclusion (Le Loarer and Normier, 1996; Monteil, 1995) to extract its topic, or, automatically by introducing some keywords to a machine (Abdelkader, 2002) using statistic or linguistic tools in 1957, (Patrice, 2000) proposes to use the frequency of words to index documents. Maron and Kuhns (1960) create the first system of automatic indexing of documents, KWIC, based on notions of probabilistic indexing. In 1966, Salton (Anderberg, 1973) followed these precursors closely and put in evidence the necessity of the automatic indexing.

Corresponding Author: Faten Khalfallah Hammouda, Department of Informatics, Faculty of Economy and Management, Sfax, Tunisia

In this study, we introduce the whole system, System of Indexation and Classification of Arabic Texts (SICAT) permitting the indexation and classification of Arabic texts using a heuristic approach. This approach is based on calculating the number of apparition of different keywords in the text to classify. A similar work was done two years ago by (Rebai and Ben Ayed, 2007), but it was not achieved for all domains. It was based on a linguistic approach and the results for the domain of medicine were approximately the same as the obtained results in our system.

The rest of the study is organized as follows. A discussion about the system architecture (SICAT) followed by the obtained results illustrated by some experimentation to validate it. Finally, we conclude the study and point out the future work.

System architecture overview: The entire architecture of System of Indexation and Classification of Arabic Texts (SICAT) is presented in Fig. 1 mainly the system can be classified into two parts: Indexation and classification. The indexation includes the pre-learning, calculating the number of words before lemmatization, the lemmatization and the calculation of frequencies. On the other hand, in the classification stage, we extract the different keywords from the text and compare them with the list of keywords extracted in the pre-learning stage to classify finally the text in the suitable domain.

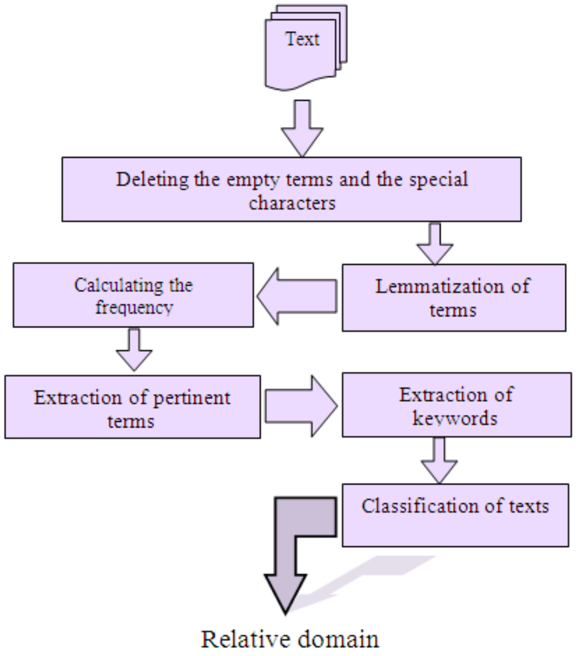

Text
Deleting the empty terms and the special characters
Lemmatization of terms
Calculating the frequency
Extraction of pertinent terms
Extraction of keywords
Classification of texts
Relative domain

Fig. 1: The system architecture overview

Indexation: The main component of the system is the indexation. It is composed of several stages permitting finally the operation of indexation. First, it requires a pre-learning step in which we must constitute a corpus and study the different texts to extract the pertinent (important) terms. Second, we move to the lemmatization step where we'll describe how to modify the different words having the same root to a unique form. Finally, we'll present the different formula to calculate the words' frequencies using the TFIDF approach.

Pre-learning: The pre-learning stage is the most important in the architecture of the proposed system SICAT to obtain the better results. It consists on collecting a corpus of Arabic texts from different domains: Sports, medicine, agriculture, informatics and politics. Besides, to study them by extracting the different pertinent terms for every domain, that means, the words which frequently used for every domain.

Constitution of corpus: For the part concerning the training and in order to get a sufficiently objective word list, we extracted 25 texts (5 texts by domain) from Internet, newspapers and magazines to form manually a corpus composed of around 12500 words. After extracting these texts, we made some modifications on the preposition of conjunction “و ” because this last can't be distinguished in the Arabic language if it is about a preposition or a letter belonging to a word because this letter is always attached to the following letter. That is why we distinguished the preposition from the letter by putting manually a space around it.

Constitution of pertinent terms: During this phase we construct a database from the elaborated corpus containing different texts belonging to the five domains knew. This database contains the pertinent terms related to every domain. Indeed, a document contains generally an important number of special characters (#, %, @, *, /, +, ? and !), of characters bringing little information on the text itself and to disable the training. Therefore, we go at this stage “to filter” or “to clean” the text in order to only preserve the necessary information to the training.

Thus, all texts treated will be formatted with the same way. Besides, we know that every term have many flexional forms. In order to be able to calculate the frequency of words in a text, all terms must be lemmatized. So, it is about rewriting every text with a simpler alphabet.

We'll start by deleting the non alphabetic characters and the characters that don't belong to the

Arabic language (1,2,3,4,5,6,7,8,9,0,٢,٣,٧,٩,٨,٥,٤, ١,٦,&, », «, ‘, -,’, ., ^,@,],[,(,),{,},/,\,§,£,*,-,A,B,C,D,…, X,Y,Z,a,b,c,d,e,f, …, y, z). After having finished this treatment, we’ll have a text composed only of Arabic words. On this text we’ll apply some modifications. We’ll eliminate the particle “كالـ ، فالـ ، بالـ ، للـ” situated on the beginning of the word, the particles «“الـ” situated on the beginning of the word, the letters ” ه ، ـه ، ة ، ـة ”situated at the end of words, the empty words (من ، إلى ، عن ، على ، ف), the words having less than 3 letters and the words situated in the dictionary of empty words.

Generally, the empty words represent some linguistic tools used to link or coordinate the sentences. Some empty words, like the words “قديما”, “تقريبا” … rarely appear in texts. By calculating the discrimination value, we don’t arrive to eliminate them. But, we don’t like to present them as index because they do not have a sense.

To eliminate these words, we use an anti-dictionary which contains a list of the words that we don’t like to keep them. These words are often prepositions (“ من “,” إلى “), pronouns (“لا”, “كل”, “هو” ,“أنا”), adverbs (“الآن”, “حتى”), adjectives (“بعيد”, “ممكن”).

In conclusion, the process of filtering the empty words consists in the suppression of words that make the documents similar but its content of information is different.

Lemmatization: In this stage, we concentrate on the lemmatization step of some terms. This phase has as goal to make all the terms, which have the same root, in a unique form to facilitate the treatment. We can take as an example the terms: زرع ، مزارع ، زراعات ، مزرعة in a text. When we ask for the operation of lemmatization, all those will be transformed to زرع to calculate the frequency with an efficient way. In fact, our method for indexation and classification of Arabic texts system development is heuristic, that’s means we don’t use any linguistic rules. The system analyzes all the formatted text; it makes a comparison by two words. That’s means that it will look for the first word in the second one and vice-versa.

This research is based on the position of the letter, situated in the first word, in the second one. These positions must be classified in an increasing order because we can find the same letters in two words semantically different. If this condition has been verified, the longest word will be replaced by the smallest. The Table 1 will illustrate. The first example shows that the positioning order of the first word letters in the second one and vice versa is increasing (1, 2, 3 and 2, 3, 4), that’s why the longest will be transformed to the smallest, it means مزرعة will be زرع without any linguistic recourse. On the contrary, in the second example, the positioning of letters of the first word in the second one and the second in the first is in mess (3, 1, 2 and 3, 1, 2), therefore these two words are considered as different and no modification will be done on one of them.

Calculating the word apparition frequency: This stage permit to determine the value of frequency of every term in every domain to be able to extract the pertinent terms after having lemmatized all the text’s terms. In this study, we’ve based on the TFIDF approach to determine the pertinent terms. With Term Frequency (TF), we designate a measure that has report to the importance of a term for a document. In general, this value is determined by the frequency of the term in the document. By Inverted Term Frequency (ITF), we measure if the term is discriminative (or doesn’t uniformly distributed). Here, we give the formulas of TF and IDF used for our system:

- TF = log (f (t, d) + 1) with f (t, d) is the number of occurrence of the term t in the documented
- IDF = log (N/n) With N is the number of documents in the corpus, n is the number of documents that contain the term
- The formula TFIDF combines the two criterions that we’ve seen
- The importance of a term for a document (by TF) and the power of discrimination of this term (by IDF)
- The multiplication of TF by IDF:

$$TF*IDF = TFw,d * \log(N/n)$$

Thus, a term that has a high TF*IDF value must be simultaneously important in this document and it must appear little in the other documents. It is the case where a term correlates to an important and unique characteristic of a document. With such formula, we can choose to keep only terms that the value of TFIDF exceeds a doorstep. These terms are called the pertinent terms and this doorstep is the average of frequencies.

Table 1: Illustration of heuristic lemmatization method

	First word				Second word				
1st example	ـة	عـ	ر	ـز	مـ		ع	ر	ز
Position in the other word	0	3	2	1	0		4	3	2
2nd example			ـب	ـعـ	لـ	ـة	ـبـ	ـلـ	عـ
Position in the other word			3	1	2	0	3	1	2

Extraction of pertinent terms: After calculating the frequency of all the terms in every domain, we'll extract the pertinent terms. This extraction is done using a simple arithmetic formula. This method is composed of two stages: Calculating the average of frequencies for all the terms using the following formula:

Frequencies average = Total of TFIDF/Number of terms and comparing the frequency of every term with the frequencies average

If its value is superior to the average, so this term is pertinent, else it is not.

Classification: The main goal of the system (SICAT) is the classification of a text in the suitable domain from five: sports, medicine, politics, informatics and agriculture. In this stage, we'll extract the different keywords and calculate the number of its apparition. And according to these numbers, the system will classify the text in the available domain.

Extraction of keywords: Once the pertinent terms are extracted, it is about determining the groups of terms associated to every domain. At this stage, the pertinent terms of a domain will be regrouped in different wholes. Thus, a pertinent term group is frequently appearing in the same sentences.

While applying all previous treatments, we arrange for every domain, a list of pertinent terms which are lemmatized. Finally, in the last stage we determine the doorstep of every domain.

A new text classification: This stage is the last one in our system and it represents its main objective; it permits us to classify finally an unknown text in its suitable domain. So, it's to detect the text theme. To do this task, we applied a method by pertinent terms correspondence. It is about testing the correspondence of all pertinent terms of the text to classify with the keywords of every domain of the corpus. The domain, that constitutes the majority of terms having a correspondence with terms of the text, represents the theme that we look for to classify our unknown text.

MATERIALS AND METHODS

Mathematical basis of SICAT: TFIDF is the most common weighting method used to describe documents in the Vector Space Model, particularly in IR problems. The TFIDF function weights each vector component (each of them relating to a word of the vocabulary) of each document on the following basis. First, it incorporates the word frequency in the document. Thus, the more a word appears in a document (e.g., its TF, term frequency is high) the more it is estimated to be significant in this document. In addition, IDF measures how infrequent a word is in the collection. This value is estimated using the whole training text collection at hand. Accordingly, if a word is very frequent in the text collection, it is not considered to be particularly representative of this document (since it occurs in most documents; for instance, stop words). In contrast, if the word is infrequent in the text collection, it is believed to be very relevant for the document. TFIDF is commonly used in IR to compare a query vector with a document vector using a similarity or distance function such as the cosine similarity function. There are many variants of TFIDF. The following common variant was used in our experiments, as found in (Yang and Liu, 1999).

$$\text{weight}_{t,d} = \begin{cases} \log(\text{tf}_{t,d} + 1)\log\dfrac{n}{x_t} & \text{if } \text{tf}_{t,d} \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

Where:

$\text{Tf}_{t,d}$ = The frequency of word t in document d
n = The number of documents in the text collection
x_t = The number of documents where word t occurs

Used environment: In this study, we have used the visual basic platform. This platform allowed us to create this application permitting to classify Arabic texts in one of the chosen domains: sports, medicine, economics, agricultures and politics. In this platform, we didn't use any grammatical rule, we have interested in statistic method based on arithmetic rules and calculations. We have used the different tools of visual basic to ensure a very nice user interface easy to use and containing clear options (captions, textboxes and icons).

RESULTS

In order to support our system, we introduce some examples to illustrate it. We start by treating the stage of formatting the text. After that, we apply the lemmatization stage on the formatted text.

The origin text:

اذا اخذنا تربة العراق التي تعد واحدة من أخصب بقاع العالم في انتاج
أنواع كثيرة من المحاصيل الزراعية، معياراً للنشاط الزراعي في
العراق، فان ذلك يشير بوضوح الى ان الزراعة عندنا أصبحت من
القطاعات التي شهدت تدهوراً في مستوياتها هذا التدهور لمسه المواطن

العراقي بافتقاد منتجات بلاده الزراعية في السوق و بغلاء أسعار الفواكه
و الخضر و بقية المنتوجات الزراعية ، كما عاشته الدوائر الزراعية
نفسها وعلى رأسها وزارة الزراعة.

The text after formatting:

اخذن ترب عراق تعد أخصب بقاع عالم إنتاج أنواع كثير محاصيل زراع
معيار نشاط زراع عراق يشير بوضوح زراع قطاعات شهدت تدهور
مستويات تدهور مواطن عراق بافتقاد منتجات بلاد زراع سوق بغلاء
أسعار فواك خضر منتوجات زراع عاشت دوائر زراع رأس وزار زراع.

After having formatted the text the number of words decreased. It was 70 words and becomes 43 words. All the empty words and some particles will be deleted, that’s why the treatment will be easier. We move now to the stage of lemmatization.

Formatted text before lemmatization:

اخذن ترب عراق تعد أخصب بقاع عالم إنتاج أنواع كثيرمحاصيل زراع
معيار نشاط زراع عراق يشير بوضوح زراع قطاعات شهدت تدهور
مستويات تدهور مواطن عراق بافتقاد منتجات بلاد زراع سوق بغلاء
أسعار فواك خضر منتوجات زراع عاشت دوائر زراع رأس وزار زراع.

Text after lemmatization:

اخذن ترب عراق تعد أخصب بقاع عالم إنتاج أنواع كثير محاصيل زراع
معيار نشاط زراع عراق يشير بوضوح زراع قطاعات شهدت تدهور
مستويات تدهور مواطن عراق بافتقاد إنتاج بلاد زراع سوق بغلاء أسعار
فواك خضرانتاج زراع عاشت دوائر زراع رأس وزار زراع.

This stage of lemmatization represents a big importance for the following stage; it is the calculation of frequencies. For this reason, the stage of lemmatization proves to be useful. For a goal of illustration, we take the example in the Table 2.

Here, the average of frequencies is equal to 0.3351. In fact, if the phase of lemmatization won't be done, then the terms أراضي ، زراع will not be considered as pertinent terms because their frequency of apparition doesn't exceed the average of frequencies which represents the minimum doorstep. It’s necessary to note that the frequency must be superior to the minimum doorstep and inferior to the maximum doorstep. On the contrary, if the lemmatization is done, a lot of modifications will be treated. This will be resumed in the Table 3.

From this example, we note the importance of the lemmatization which permits to facilitate the extraction of pertinent terms. As an example, we test the correspondence of all terms of a text having as theme «AGRICULTURE» as show in the Table 4.

Table 2: Lemmatization representation

Word	Number of occurrences	Frequency
مزارع	3	4444 1.2
زراع	1	0.33333
أرض	6	1.77777
أراض	1	0.33333

Table 3: Modifications

Word	Number of occurrences	Frequency
ماء	1	0.33333
زراع	4	1.57777
أرض	7	2.1111

Table 4: Extraction

Correspondence	Terms
SPORT	1
AGRICULTURE	15
POLITICS	3
INFORMATICS	0
MEDICINE	2
RESULT	AGRICULTURE

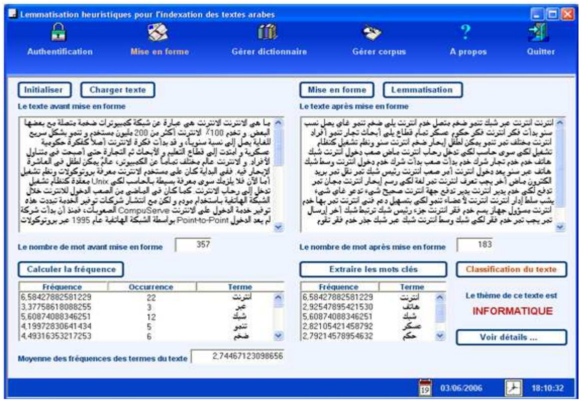
In this text, 15 terms appear as keywords for the theme «AGRICULTURE» and we find only 3 terms representing keywords for the «POLITICS» domain. So, as a conclusion, we can deduct that the text must be classified between those having the theme «AGRICULTURE». This method isn’t effective when the text to classify doesn’t contain a sufficient number of terms related to the text domain. In this case, we obtain very near values, sometimes equal or hopeless, that stops us from detecting the searched term. We can take for example these cases:

- The term ” حكم ” is a keyword for the domains «POLITICS» and «SPORTS»
- The term ” غذاء ” is a keyword for the domains «MEDECINE» and « SPORTS»
- The term ” دواء ” is a keyword for the domain «AGRICULTURE» and «MEDECINE»

DISCUSSION

The interface user machine is an interface that a user doesn’t notice anymore. We aim in our application the conversational treatment and the ergonomics adopting the updated technologies: Minimal seizure, color and questioning. Figure 2 shows some interfaces of our application and represents the different stages of our method. It classified the present text in the domain of informatics.

Figure 3 shows the interface permitting to add, remove and update entries to the dictionary. This stage is done after analyzing a text.



Le texte avant mise en forme

ما هي الانترنت الانترنت هي عبارة عن شبكة كمبيوترات ضخمة متصلة مع بعضها
البعض. و تخدم 100٪ الانترنت أكثر من 200 مليون مستخدم و تنمو بشكل سريع
للغاية يصل إلى نسبة سنوياً، و قد بدأت فكرة الانترنت أصلاً كفكرة حكومية
عسكرية و امتدت إلى قطاع التعليم و الأبحاث ثم التجارة حتى أصبحت في متناول
الأفراد. و الانترنت عالم مختلف تماماً عن الكمبيوتر، عالمٌ يمكن لطفل في العاشرة
الإبحار فيه. ففي البداية كان على مستخدم الانترنت معرفة بروتوكولات ونظم تشغيل
أما الآن فلا يلزمك سوى معرفة بسيطة بالحاسب لكي Unix معقدة كنظام تشغيل
تدخل إلى رحاب الانترنت. كما كان في الماضي من الصعب الدخول للانترنت خلال
الشبكة الهاتفية باستخدام مودم و لكن مع انتشار شركات توفير الخدمة تبددت هذه
توفير خدمة الدخول على الانترنت CompuServe الصعوبات، فمنذ أن بدأت شركة
لم يعد الدخول Point-to-Point بواسطة الشبكة الهاتفية عام 1995 عبر بروتوكولات

Le texte après mise en forme

انترنت انترنت عبر شبك تنمو ضخم متصل خدم انترنت يلي ضخم تنمو غاي يصل نسب
سنو بدأت فكر انترنت فكر حكوم عسكر تمام قطاع يلي أبحاث تجار تنمو أفراد
انترنت مختلف تمر تنمو يمكن لطفل إبحار ضخم انترنت سنو ونظم تشغيل كنظام
تشغيل لكي سوى حاسب لكي تدخل رحاب انترنت ماض صعب دخول انترنت شبك
هاتف خدم خدم تجار شرك خدم بدأت صعب بدأت شرك خدم دخول انترنت وسط شبك
هاتف عبر سنو يعد دخول انترنت أمر صعب انترنت رئيس شبك تمر نقل تمر بريد
الكترون ماض أخر يجب تعرف انترنت تمر لغة لكي رسم إبحار انترنت مجان تمر
تدفع لكي خدم يدير انترنت يدير تدفع جهة انترنت صحيح شيء تدعو غاي شيء
يشب سلط إدار انترنت انترنت لأعضاء تنمو لكي بتسهيل دعم فني انترنت تمر بها خدم
انترنت مسؤول جهاز يسم خدم فقر انترنت جزء رئيس شبك ترتبط شبك أخر إرسال
تمر يجب تمر خدم فقر لكي شبك وسط انترنت شبك عبر شبك جذر خدم فقر تقوم

Le nombre de mot avant mise en forme 357

Le nombre de mot après mise en forme 183

Fréquence	Occurrence	Terme
6,58427882581229	22	انترنت
3,37758618088255	3	عبر
5,60874088346251	12	شبك
4,19972830641434	5	تنمو
4,49316353217253	6	ضخم

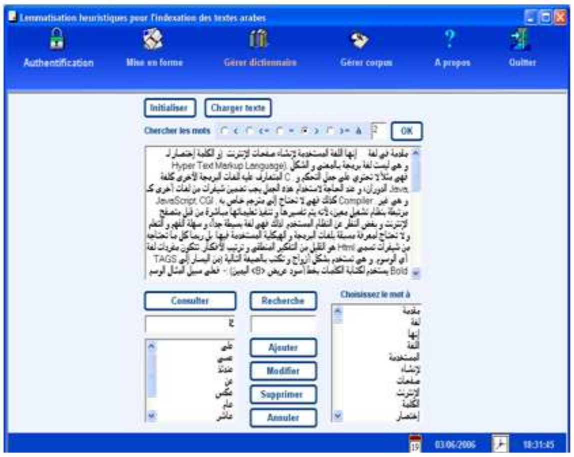
Fréquence	Terme
6,58427882581229	انترنت
2,92547895421530	هاتف
5,60874088346251	شبك
2,82105421458792	عسكر
2,79214578954632	حكم

Le thème de ce texte est INFORMATIQUE

Moyenne des fréquences des termes du texte 2,74467123098656

03/06/2006 18:10:32

Fig. 2: Different stages of the system



Chercher les mots ○ < ○ <= ○ = ● > ○ >= à 2

مقدمة في لغة   إنها اللغة المستخدمة لإنشاء صفحات لإنترنت و الكلمة إختصار ل
و هي ليست لغة برمجة بالمعنى و الشكل Hyper Text Markup Language).
فهي مثلاً لا تحتوي على جمل التحكم و C المتعارف عليه لغات البرمجة الأخرى كلغة
Java. الدوران، و عند الحاجة لاستخدام هذه الجمل يجب تضمين شيفرات من لغات أخرى ك
و هي غير Compiler كذلك فهي لا تحتاج إلى مترجم خاص به . JavaScript, CGI
مرتبطة بنظام تشغيل معين، لأنه يتم تفسيرها و تنفذ تعليماتها مباشرة من قبل متصفح
لإنترنت و بغض النظر عن النظام المستخدم. لذلك فهي لغة بسيطة جداً، و سهلة الفهم و التعلم
و لا تحتاج لمعرفة مسبقة بلغات البرمجة و الهيكلية المستخدمة فيها بل ربما كل ما تحتاجه
من شيفرات تسمى Html هو القليل من التفكير المنطقي و ترتيب الأفكار. تتكون مفردات لغة
أي الوسوم، و هي تستخدم بشكل أزواج و تكتب بالصيغة التالية (من اليسار إلى TAGS
Bold يستخدم لكتابة الكلمات بخط أسود عريض <B> اليمين) - فعلى سبيل المثال الوسم

ع

على
عسى
عندئذ
عن
عكس
علم
عاشر

Choisissez le mot à

مقدمة
لغة
إنها
اللغة
المستخدمة
لإنشاء
صفحات
لإنترنت
الكلمة
إختصار

03/06/2006 18:31:45

Fig. 3: Updating dictionary

Such as the majority of the indexation applications, the two measures would be used in this application are the recall and the precision. The recall measures the proportion of extracted pertinent texts in report with all the pertinent texts and the precision measures the proportion of pertinent texts in report with all the extracted texts. To use these two measures, it's necessary to be able to determine the pertinent texts for every request. The evaluation of the system performance on the different chosen domains, achieves 90% precision and 85% recall.

Due to this work, we could classify a big number of Arabic texts in the suitable domain. We could also elaborate specific dictionaries related to the chosen domains (sport, economy, medicine, politic and agriculture). The output of this study can be used in other NLP applications such as the information extraction and automatic abstract. This application can accept any update in its dictionaries or its domains.

CONCLUSION

In this study, we have described our System for Indexation and Classification of Arabic Texts (SICAT). It holds two main parts: The indexation and the classification. The indexation stage is composed of three main parts: The pre-learning, the lemmatization and the frequencies' calculation. The classification stage is composed of two main components: the extraction of keywords and classification of new text. We have made many tests of verification to test the validation of the system. The system performance was evaluated on the different chosen domains, achieves 90% precision and 85% recall.

We have considered only some simple words as representatives of concepts which are also called indexes. In the future, we plan to increase the number of domains to be able to classify more texts in more themes.

REFERENCES

Abdelkader, H., 2002. Master, Markovian approach for thematic classification of Arabic texts. pp: 106-112.

Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York, ISBN: 10: 0120576503, pp: 359.

Le Loarer, P. and Normier, 1996. Linguistic and statistical techniques to select relevant information. Proceedings of the Conference on IDT'96, pp: 115-120.

Maron, M.E. and J. Kuhns, 1960. On relevance, probabilistic indexing and information retrieval. J. Assoc. Comput. Mach., 7: 216-244.

Monteil, M.G., 1995. Manual indexing and automatic indexing: Comparison and prospects. Proceedings of the 12th Congress on IDT'95, Paris, pp: 210-214.

Patrice, B., 2000. Methods for Classification and Segmentation for local unsupervised retrieval. Thesis, pp: 206-210.

Rebai, M. and M. Ben Ayed, 2007. Clatexa, Design and implementation of a classification system of Arabic texts.

Yang, Y. and X. Liu, 1999. A re-examination of text categorization methods. Proceedings of the 22nd annual International ACM SIGIR Conference on Research and development in Information Retrieval, Aug. 15-19, ACM Press, Berkeley, California, United States, pp: 42-49. http://portal.acm.org/citation.cfm?id=312647