

# Semantic internet search engine with focus on Arabic language

Naima Tazit, El Hossin Bouyakhf, Souad Sabri

*Faculty of Sciences Agdal, Rabat, Morocco*

Abdellah Yousfi

*Faculty of Economy Souissi, Rabat, Morocco*

Karim Bouzouba

*Mohammadia School of Engineers, Rabat, Morocco*

**Keywords :** Search engine, semantic, Arabic, Harman, Croft, Okapi.

**Abstract** We present in this paper an Internet search engine with focus on the Arabic language. To develop such a search engine, we used regular document retrieval techniques and enhance them with a treatment on the semantic level of terms found in documents. This semantic process is integrated in the search stage of the search. We evaluated the developed system using several weight functions (Harman, Croft, and Okapi).

21

## Introduction

Today the web includes many Arabic portals (Aljazeera, an Arabic Television Channel, advertises more than seven million (7,000,000) documents in its web site [www.aljazeera.net](http://www.aljazeera.net), etc.) and several search engines [1, 2, 3, 4]. In general, for any human language the market for sophisticated search engines is still growing, and more so for the Arabic Internet.

The inefficiencies in retuning pertinent documents (inefficiencies in precision and recall, and in response time [5]) are hindering the use of this fast-growing library that is the Arabic Internet. It is therefore of extreme importance to the development of Arabic content on the web that more efficient techniques and systems be developed. We believe that by incorporating techniques from Artificial Intelligence and from Arabic word computational morphology and semantics, many of the limitations could be overcome.

Indeed, the quality of a search engine depends not only on precision and response time but it depends also on linguistic processes it may include. Kinds of problems we may encounter when using search engines are: morphological variation and lack for considering the semantic level of the language. However, current Internet search engines don't take into consideration at the same both the morphological and the semantic level of the language when searching for a specific word.

To develop search engines, document retrieval techniques are used. The aim of information retrieval or document retrieval is to develop computing tools that allow users to find information that is located in a great number of documents [6]. The user can ask for information in two ways:

- Individual (isolated, selective): the user ask the system using queries
- Persistent: the retrieval is known in advance

In this work, we present an Arabic search engine that integrates a semantic process. For this purpose, the search engine uses terminological dictionaries of Arabic words.

In general, the process for document retrieval is done in three steps:

1. Indexing documents
2. Searching from a user request
3. Presenting the results of the search

In this paper, we will mainly focus on the semantic level of the language. We will briefly start presenting the indexing step. Then, we will discuss how the similarity measure is used to develop the semantic level of the Arabic language and how it is used in the search engine. Finally, we will present some results of our experiments.

## Indexing:

Indexing is at the heart of any search engine. Our model does not change a lot from the existing indexing methods (remove of empty words, etc.). It consists of representing every document belonging to a Collection C with an array of terms. To each term, we associate a set of attributes



(position of the term in the collection, the weight of the term in every document). In our approach, we rely on the three most used weight functions (Harman, Croft and Okapi) to reflect the importance given to a document. Let:

- C: be a collection of documents,  $C = \{d_1, d_2, \dots, d_N\}$
- V: be the set of all the existing terms in the index of C.

**The search procedure:**

The stage of search takes place through a request in which the user formulates his demand. Among the approaches used in this stage, we can mention [7]:

- Boolean approach: the inconvenience of this approach is that it doesn't take into account the notion of relevance of documents [8].

- The approach by similarity measure: this approach allows assigning to every document a score of relevance [9].

If we note by  $Q = \{x_1, x_2, \dots, x_p\}$  a request, the score or the similarity value of a document d concerning Q is:

$$S(Q, d) = \frac{\sum_{i=1}^p F(x_i, d)}{\sum_{i=1}^p P_C(x_i)}$$

F(x,d): is a function that represents the weight of the term x in document d. In our system the value of this function is calculated from one of the three known weight functions (Herman, Okapi, and Croft functions). Details that explain how our system is using these functions are outlined in the next section.

The set of documents that answer the request Q is given by the following search function R [9]:

$$R(Q) = \{ (d_i, a_i) / d_i \in C \text{ and } a_i = S(Q, d_i) > 0 \}$$

Documents R(Q) are ordered using the score S(Q, d).

**Use of semantic:**

Current search engines return documents that contain only the word w to search. However, some other documents may be of interest because they contain semantically close words to w.

Example:

2	1
يتكون الطب من عدة مجالات مهمة، نذكر منها: علم التشريح، علم الأوبئة وانتقال الفيروسات. كما أنه يختص بدراسة الأمراض ومحاولة إيجاد أدوية فعالة.	يعتبر الطب من المجالات المهمة التي تحظى باهتمام الدولة، لكن رغم هذا الاهتمام يظل الطب يحتاج إلى المزيد من الدعم المادي.

For current search engines, searching the word “الطب” (Medical Science) using only the existence of that word in documents, document1 is more relevant than document2. However, document2 is richer in terms of words belonging to the medical domain “التشريح، الأوبئة، الأمراض، أدوية” (Medicines, Diseases, epidemics, surgery). Hence, developing a search engine taking into account the semantic level of words would be very interesting since it may help users retrieving documents with more relevance.

In order to use the semantic level of a language during the search step, we developed terminological dictionaries. To each term  $x_i \in V$  is associated in the dictionary:

$$Dic(x_i) = \{t_{i_1}, t_{i_2}, \dots, t_{i_k}\}$$

We replaced the search function for the term  $x_i$  with the following function:  $R(x_i, t_{i_1}^\theta, t_{i_2}^\theta, \dots, t_{i_k}^\theta)$   
 $\theta$  : represents the importance given to terms  $\{t_{i_1}, t_{i_2}, \dots, t_{i_k}\}$  for  $x_i$  in the search function. In this case, the similarity measure is given as follows:

$$S((x_i, t_{i_1}^\theta, t_{i_2}^\theta, \dots, t_{i_k}^\theta), d) = \frac{F(x_i) + \theta \times \sum_{j=1}^k F(t_{i_j}, d)}{P_C(x_i) + \sum_{j=1}^k P_C(t_{i_j})}$$

In order to evaluate our model, we used the following three similarity measures:

- **Harman measure** : the  $F(x, d)$  has the following value [10] :

$$F(x, d) = p_c(x) \times f_d(x)$$

With:

$$p_c(x) = -\log(n(x) / N) \text{ and } f_d(x) = \frac{\log(o_d(x) + 1)}{\log(L_d)}$$

- $n(x)$  : the number of documents containing  $x$ .
- $N$ : number of documents in collection  $C$ .
- $O_d$  : is the occurrences number of  $x$  in document  $d$ .
- $L_d$  : the size of  $d$ .

- **Croft measure**: the  $F(x, d)$  has the same value that Harman, with:

$$f_d(x) = k + (1 - k) * o_d(x)$$

$K$  is a constant of importance of  $x$  in  $d$ . The best results are given by  $K=0,5$  [11].

- **Okapi measure**:  $F(x, d)$  has the same value that Harman, with:

$$p_c(x) = \log\left(\frac{N - n(x) + 0.5}{n(x) + 0.5}\right) \quad f_d(x) = \frac{o_d(x)}{o_d(x) + \frac{L_d}{L}}$$

$L$ : the middle size of a document of  $C$ .

### Experiments and results:

To evaluate our approach, we used a collection of 100 documents and four terminological dictionaries. The index of this collection is built in such a way that each term of this index is associated to a set of attributes (width of the term, position of the term ...).

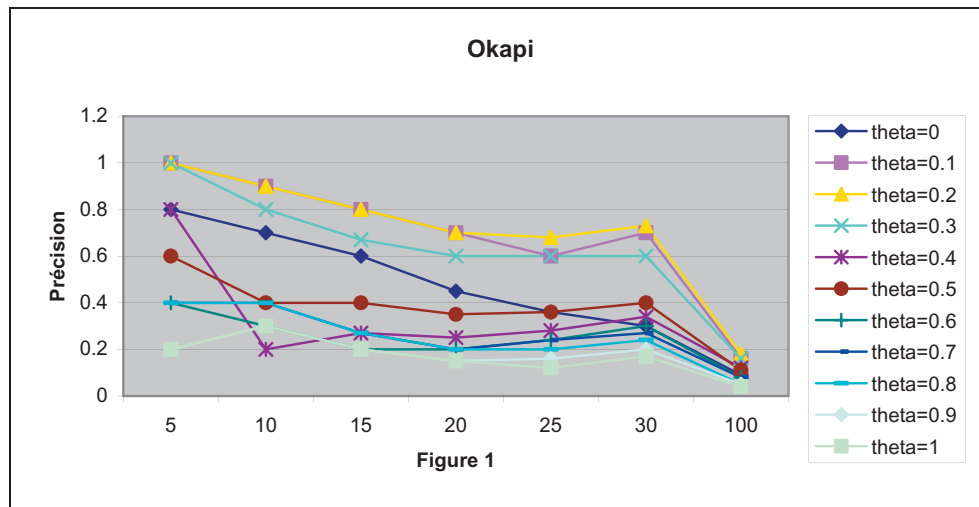
We then varied the value of  $\theta$  for each of the three similarity measures listed above and calculated the precision for every value of  $\theta$  and every measure (table1, table2, table3)

Okapi measure:

**Table 1.**  
Evolution of the precision according to theta by using Okapi measure

$\theta$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
N=5	80%	100%	100%	100%	80%	60%	40%	40%	40%	20%	20%
N=10	70%	90%	90%	80%	20%	40%	30%	40%	40%	30%	30%
N=15	60%	80%	80%	67%	27%	40%	20%	27%	27%	20%	20%
N=20	45%	70%	70%	60%	25%	35%	20%	20%	20%	15%	15%
N=25	36%	60%	68%	60%	28%	36%	24%	24%	20%	16%	12%
N=30	30%	70%	73%	60%	34%	40%	30%	27%	24%	20%	17%
N=100	9%	20%	22%	18%	10%	12%	9%	8%	7%	6%	5%

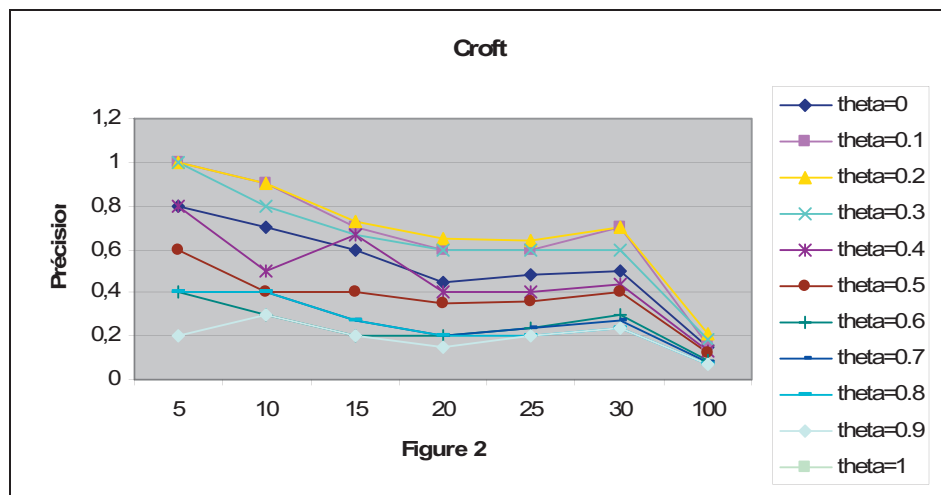
24



Croft measure:

**Table 2.**  
Evolution of the precision according to theta by using Croft measure

$\theta$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
N=5	80%	100%	100%	100%	80%	60%	40%	40%	40%	20%	20%
N=10	70%	90%	90%	80%	50%	40%	30%	40%	40%	30%	20%
N=15	60%	70%	73%	67%	47%	40%	20%	27%	27%	20%	14%
N=20	45%	60%	65%	60%	40%	35%	20%	20%	20%	15%	10%
N=25	48%	60%	64%	60%	40%	36%	24%	24%	20%	20%	16%
N=30	50%	70%	70%	60%	44%	40%	30%	27%	24%	24%	20%
N=100	15%	16%	21%	18%	13%	12%	9%	8%	7%	7%	6%



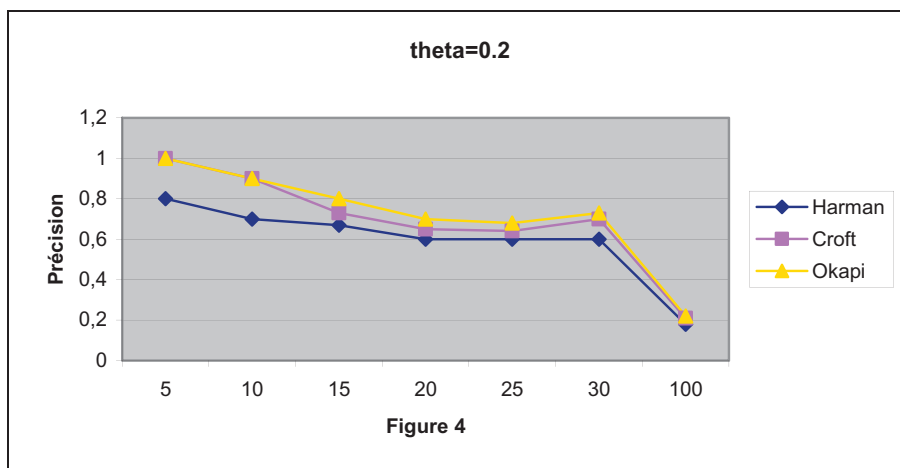
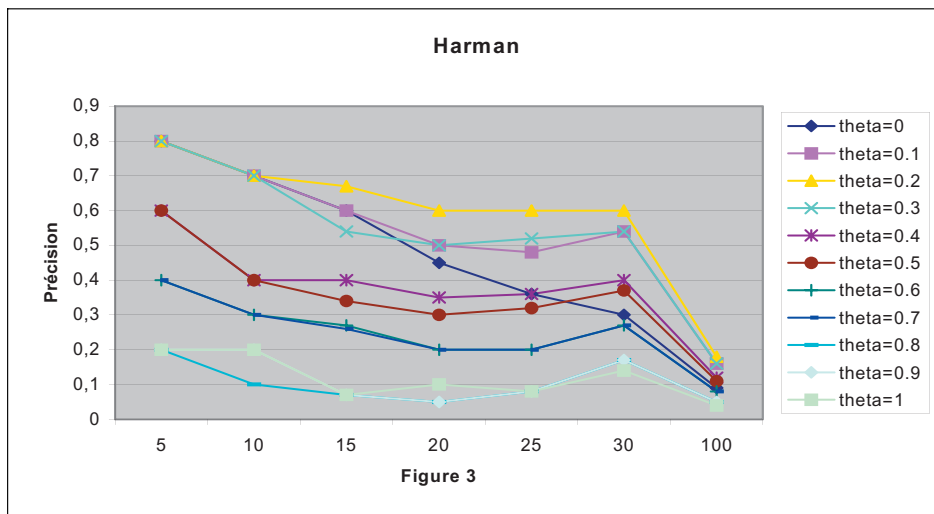
Harman measure:

$\theta$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
N=5	80%	80%	80%	80%	60%	60%	40%	40%	20%	20%	20%
N=10	70%	70%	70%	70%	40%	40%	30%	30%	10%	20%	20%
N=15	60%	60%	67%	54%	40%	34%	27%	26%	7%	7%	7%
N=20	45%	50%	60%	50%	35%	30%	20%	20%	5%	5%	10%
N=25	36%	48%	60%	52%	36%	32%	20%	20%	8%	8%	8%
N=30	30%	54%	60%	54%	40%	37%	27%	27%	17%	17%	14%
N=100	9%	16%	18%	16%	12%	11%	8%	8%	5%	5%	4%

Semantic internet search engine

25

**Table 3.** Evolution of the precision according to theta by using Harman measure



**Figure 1.** Comparison of three measures (Harman, Croft, Okapi)

The graphs 1, 2, 3 show that the best results are got for  $\theta = 0.2$ . The graph 4 shows that the number of relevant documents returned by Okapi measure is greater than to the one of Croft and Harman.

### Conclusion and perspectives

We presented in this work an Internet search engine with focus on the Arabic language. The developed search engine integrates regular document retrieval techniques with an enhancement on the semantic level of terms found in documents. We evaluated the developed system using several weight functions (Harman, Croft, Okapi).

The best results are given by the Okapi measure and for  $\theta = 0.2$ . In this work we supposed that all terms of dictionaries have the same relevance  $\theta$ . It would be interesting to calculate a specific relevance for every term in the dictionary. Then, these dictionaries (with specific relevance for each term) have to be used in the indexing stage instead of the search stage.

## References

- [1] Google, <http://www.google.com/>
- [2] Hahooa, <http://www.bahooa.com/nav.php?ver=ar>
- [3] Konouz, <http://www.konouz.com>
- [4] Maktoob, <http://www.maktoob.com>
- [5] T. Rachidi et al, Barq: "Distributed multilingual Internet search engine with focus on Arabic language".
- [6] Salton G., McGill M. J., (1983) "Introduction to modern information retrieval". McGraw-HillBook Company, 448 p.
- [7] Yousfi A., (2006), "Traitement automatique de la langue, texte et parole". Edition et Impressions Bouregreg , ISBN : 9954-423-98-2.
- [8] Mooers C., (1960) "Mooers' Law or, Why Some retrieval System Are used and Others Are Not". American Documentation, Vol. 11, No. 3.
- [9] Claude D. L., "Evaluation de l'apport de connaissances linguistique en désambiguïsation sémantique et recherche documentaire". Thèse de doctorat, 2000.
- [10] Harman D., (1986) "An experimental study of factors important in document ranking". Actes de ACM, Conference on Research and development in Information Retrieval; Piste, Italie.
- [11] Salton G. and Buckley C., (1988) "Term-weighting approaches in automatic text retrieval. Information" Processing and Management, 24(5); pp. 513-523.