

Semantic Internet search engine with focus on Arabic language

Tazzit N., Yousfi A., Bouzoubaa K.,
Bouyakhaf E., Sabri S.

1. Document Retrieval

- D.R: Find information that is located in a great number of documents
- Three steps :
 - Indexing documents
 - Searching from a user request
 - Presenting the results of the search

1.1 Indexing

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 every document).

1.2 Searching

- The user formulates his demand.
- Among approaches :
 - Boolean approach: combines terms with logical expressions (AND, OR, NOT)
 - The approach by similarity measure: this approach allows assigning to every document a score of relevance

Similarity approach

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:

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

$F(x_i, d)$ is a function that represents the weight of the term x in document d .

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

The set of documents that answer the request Q is given by the following search function R.

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

2. Use of semantic

D1

يعتبر الطب من المجالات المهمة التي تحظى باهتمام الدولة، لكن رغم هذا الاهتمام يظل الطب يحتاج إلى المزيد من الدعم المادي .

D2

يتكون الطب من عدة مجالات مهمة، نذكر منها: علم التشريح ، علم الأوبئة وانتقال الفيروسات . كما أنه يختص بدراسة الأمراض ومحاولة إيجاد أدوية فعالة .

■ D2 is richer in terms of words belonging to the medical domain (Medicines, Diseases, epidemics, surgery).

■ We developed terminological dictionaries.

■ To each term $x_i \in V$ is associated in the dictionary:

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

2.1 Adaptation de la fonction de recherche

Nous avons remplacé la fonction de recherche du terme par la fonction de recherche suivante :

$$R(x_i, t_{i_1}^\theta, t_{i_2}^\theta, \dots, t_{i_k}^\theta)$$

θ : est l'importance donné aux termes $\{t_{i_1}, t_{i_2}, \dots, t_{i_k}\}$ par rapport à x_i dans la fonction de la recherche.

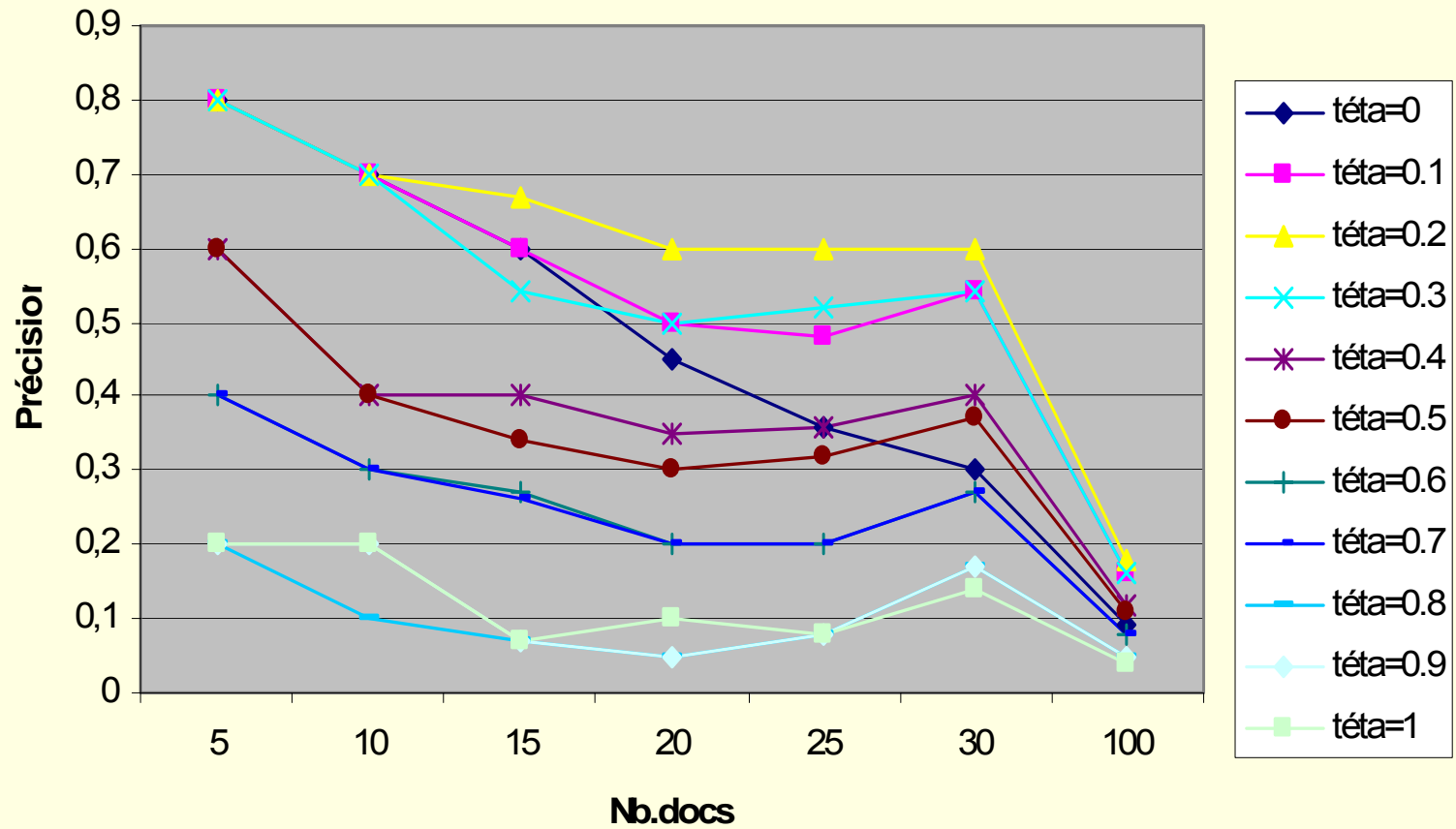
$$S((x_i, t_{i_1}^\theta, t_{i_2}^\theta, \dots, t_{i_k}^\theta), d) = \frac{F(x_i) + \theta \times \sum_{j=1}^k F(t_{i_j}, d)}{P_C(x_i) + \sum_{j=1}^k P_C(t_{i_j})}$$

3. Experiments and Results

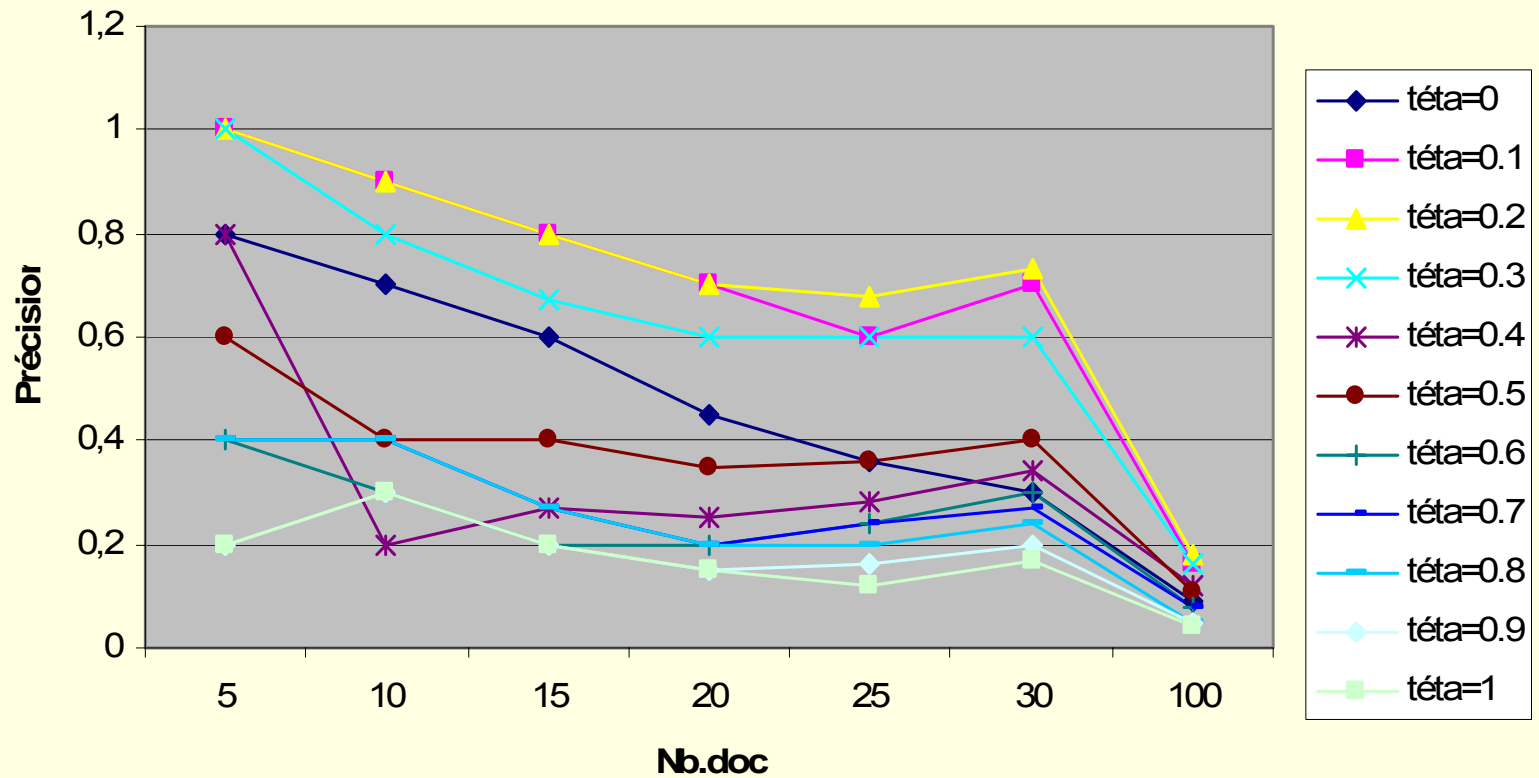
Three measures of similarity: Harman, Croft, Okapi.

We varied the value of θ for each of the three similarity measures listed above and calculated the precision θ for every value of θ and every measure

Harman

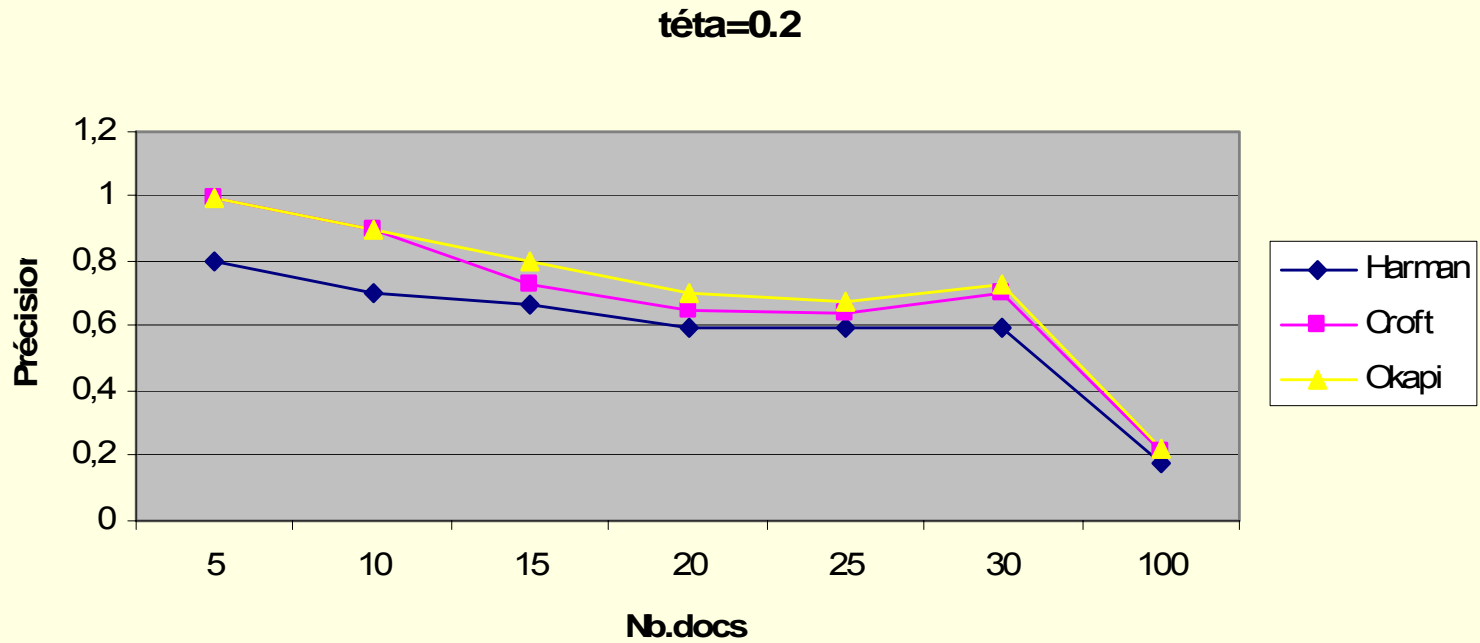


Okapi



Best results are got for $\theta = 0.2$

Comparing the three measures



4. Conclusions and perspectives

- 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. 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.