CUSTOM FUSION BASED ON BARYCENTER

[1]ABDELBAKI ISSAM, [2]BEN LAHMAR EL HABIB, [3]LABRIJI ELHOUSSIN
Department of Mathematics and Informatic, Faculty of Sciences Ben M'SIK
University HassanII – Mohammadia, Casablanca, Morocco
E-mail: [1]i.abdelbaki@gmail.com,[2]h.benlahmer@gmail.com,[3]labriji@yahoo.fr

ABSTRACT

Searching for information on the Internet is not only an activity newly rediscovered, but also a strategic tool to achieve a wide variety of information. Indeed, it's extremely important to know how to find the information quickly and efficiently. Unfortunately, the Web is so huge and so little structured, that gathering precise, fair and useful information becomes an expensive task. In order to define an information retrieval tool (meta search engine) that brings together multiple sources of information search, interest must be credited to the merger phase of search engines results. On the other hand, information search systems tend primarily to model the user with a profile and then to integrate it into the information access chain, to better meet its specific needs.

This paper presents a custom fusion method based on physical Barycenter method and values retrieved from the user profile. We evaluated our approach on multiple domains and we present some experimental results.

Keywords: *Information search system, meta-search engine, Fusion, Barycenter, User profile.*

1. INTRODUCTION

Search engines are the most visited sites on the web, they are used by 85% of users (Schwartz, 1998). However, they index only a fraction of all available information and their coverage does not increase as rapidly as the size of the Web. Thus, the user is quickly lost in finding relevant information. Meta search turns out be a powerful way to work around this problem, by bringing together multiple sources of information search (search engine) in a single unified tool (meta search engine). However, among all the problems related to the meta research, lies the fusion and the classification of the search engines results.

Having obtained an ordered list of documents from each engine, meta engines should then merge these responses in order to present a single list to the user. The response quality of meta-search depends on the classification strategy. To solve this merger problem, several works have emerged. (Selberg, 1999) proposed a strategy named "everyone has his turn", it builds the final list by taking an element of each list in the different engines by descending order. (Yager and Rybalov, 1998) suggest a policy named "everyone has his turn" giving greater importance to the lists longer than to the rank of documents. Sometimes, search engines provide a score representing the similarity degree between the request and the document. This strategy is called "fusion by score"'. However, search engines apply heterogeneous classification algorithms, so we cannot normalize the score provided by the search engine. Analysis of user's behavior reveals a particular importance. Indeed, it is by knowing perfectly how the user is going to develop its information retrieval strategies that it will be possible to propose significant information for his research. Modeling profiles and how to adapt them to different users who do not have a specific idea of the information that they are looking for, allows us to offer personalized access to scientific papers based on the user profile exploitation. There are several user profile definitions; (Wahlster et al. 1986) defined it as follows: "a user profile (or even user model) is a data set concerning the user of a computer service. It is a knowledge source that contains acquisitions on all user aspects that may be useful for the system behavior ". The user profile is extracted from the history of requests from users, our goal is to know the search engines and the documents in which there was consultation, these two elements are called the request "assets".

The Barycenter method has been used in several fields (Statistics, Physics ...), it can make a choice between several candidates. When classifying, meta-search engine has to choose between several results. Thus, we propose a merge approach based on the physical Barycenter method in the meta search engine. We use information recovered from the user model. Including the relationship between the request and documents and the relationship between the query and the search engine, thus one can have two scores, to know the documents score and search engines score relative to the query. These two scores are the main factors of our fusion method.

The first section presents the most used fusion method, the second section presents some work in personalized information retrieval, the third section presents our approach with its different axes, the fourth section presents some experimental results evaluating the performance of our approach and finally in the last section, was completed by a conclusion and an overview on our perspectives.

2. GENERAL INFORMATION SEARCH

Having obtained an ordered list of document from each engine, the meta engines must merge these responses in order to present a single list to the user. The quality of the meta engine response depends strongly on the ranking strategy.

To solve this merger problem, several works have emerged. (Selberg, 1999) proposed a strategy named "everyone has his turn", it builds the final list by taking an element of each list in the different engines by descending order. (Yager and Rybalov, 1998) suggest a policy named "everyone has his turn" giving greater importance to the lists longer than to the rank of documents. Sometimes, search engines provide a score representing the similarity degree between the request and the document. This strategy is called "fusion by score". However, search engines apply heterogeneous classification algorithms, so we cannot normalize the score provided by the search engine. WebSum (ALO Jeanne El Jed, 2005) applies new criteria to the results provided by the search engine to reclassify the pages in relevance order for the request after checking the language and information form.

The merger may also take place under the probability estimated by logistic regression (Bookstein et al., 1992) on the basis of rank and score obtained by this document (Le Calvé & Savoy 2000). Yet, (Glover et al. 2001) use a decision theory to classify the results from various search engines.

Other methods are based on scores combination. CombSUM operator introduced by (Fox & Shaw, 1994), combines scores linearly. Indeed, the different sets considered in the merge receive the same weight. The operator CombMNZ is an extension of CombSum. The documents scores that have been found by more than one system are reinforced by being multiplied by the agreements number. However, is the reasoning of the CombMNZ operator is beneficial even if the systems share a significant number of non-relevant documents? To remedy this problem, the operator CombHMEAN combines the scores by taking the harmonic average. Finally, the Borda method proves to be a conventional method in the theory of collective choice.

3. CUSTOM INFORMATION SEARCH

Implementation of customized information research systems mainly consists of two main phases: the user modeling in a pattern that is the learning phase, and the integration of this profile in one of the access to information phases. We present in this section the main approaches used in these two phases.

3.1 User profile representation

The user center of interest is represented by his application submitted to IRS (Information Retrieval System), there are several interests representation techniques to constitute the user profile. A naive interests representation is based on key words, such as the web portals case MyYahoo, InfoQuest, etc.. There are more elaborated representations to illustrate the user interests. (Gowan, 2003 and Sieg et al., 2004) represent the interests according to weighted vectors terms, and (Sieg et al., 2005 Challam et al., 2007) present them semantically following weighted concepts of general ontology, or as concepts matrices (Liu et al. 2004).

(Gowan, 2003) and (Sieg et al. , 2004) proposed a user profile model based on vectors class each of which represents a user area of interest. The centroid classes represent the user centers of interest. The semantic representation approaches exploit reference ontology to represent user interests by the weighted vectors of the ontology used. We include the concepts hierarchy of "Yahoo" or ODP as evidence most often used in this type of approach. (Challam et al. , 2007) builds the user profile on a technique of supervised documents classification deemed relevant by a similarity measure of vector with ontology concepts of the ODP. This classification allows, on multiple search sessions, to associate each ontology concept to a weight calculated by aggregating the similarity documents scores classified under this concept. The user profile will consist of all concepts with the highest weight and representing the interests of the user centers. On the other hand (Sieg et al. , 2005) exploit simultaneously user interests represented by vectors of weighted terms and "Yahoo " hierarchy concepts. The user profile will consist of contexts each formed of a representation of an adequate research concept and the representation of the research excluded concept.

A matrix representation of the user profile is adopted in (Liu et al., 2004), the matrix is constructed from the user search history to incrementally establish categories representing the user interests and associated weighted terms reflecting the interest degree of the user for each category.

3.2 User profile exploitation

Integrating user profile in the Information Retrieval process returns to operate in the reformulation and calculation of relevance score or search results ranking. (Sieg et al., 2004) offers a personalization based on queries refinement to describe a richer query translating the proper search context using a variant of the Rocchio algorithm. Indeed, the research context is represented by a pair of classes in the hierarchy of “Yahoo” categories, the first is the correct query category and similar to one of the user's interests, the second represents the category to be excluded during the search.

Other works include the user profile in the matching function query-document. (Tamine et al., 2007a) exploit interest centers in the pairing of the IR model. The value relevance of a document to a query is no longer based on the query alone but in addition to focusing on the user who submitted it.

Finally we find the personalization approaches (Challam et al, 2007.) (Ma et al, 2007.) (Liu et al, 2004) based on the search results: they are based on the combination of the initial document rank and the rank resulting from a similarity between the document and the user profile.

4. CUSTOM FUSION BARYCENTER

Our approach is based on two main phases, namely the user profiles modeling called the learning phase that feeds our knowledge base so that the ranking algorithm can be used to merge the search engines results in the classification phase. In this section we present the main lines of our approach including the building process of user profile then the Barycenter method and its use in the classification phase will be presented.

3.3 Learning phase

In this section we are presenting our user model so that the meta search engine can use it in the classification phase. In our case we need two information, namely the relationship between the query terms and documents and the relationship between the query terms and search engines. At first, the query terms are extracted, and then we save the user interaction in our knowledge base.

4.3.1 Terms extraction

To retrieve the query terms, we chose to implement a form study as follows:

- Segmentation: Find basic units corresponding to words.
  Example: You’re (We need to identify the separator, in this case «‘» is not a separator).
- Recomposition : Find compound words.
- Lexical Analysis: Bring words to a morphological base form (conjugation, gender, number).
- Stemming: is to group words that have the same origin.

Thus, for each request R, we have a list of matching terms Ti.

4.3.2 User profile construction

Based on user interactions, we recover information about the application, i.e. the query identifier, the query terms, the consulted documents and search engines associated with these documents. Indeed, when the user enters a query, it consults some documents, and search engines that gave these documents as result are deducted. These search engines and documents are called active in relation to the query.

User profiles are stored in our knowledge base so they can be used in the classification phase.

Example:

A request R contains terms (T1-T2-T3) with multiple results, the user has viewed a set of documents (D1-D2), the search engines (M1-M3-M4) gave results in these documents, so these search engines and these documents are considered assets in relation to the request R.

Specifically, each query has an identifier and has a list of weighted terms and a set of active search engines and active documents in relation to the search query.

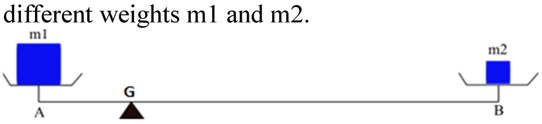
3.4 Classification phase

Our approach is a projection of the physical Barycenter method to merge results in meta search engine. As mentioned earlier, we need two information, namely the document score and the search engine score regarding the query. Scanning our knowledge base, the score represents the attendance rate each document and search engine over all query terms. In this section, we present the principle of Barycenter and then we show the utility of using this method for the search engines results classification.

4.4.1 Physical Barycenter

The Barycenter or mass center is a concept used in many scientific fields (geometry, probability, physics, chemistry ...). In Statistics it's the average notion. In mechanics it's the momentum notion and in space analysis it's the middle point or center point.

The barus Barycenter is the weight center. Its momentum and levers principles allow it to simply construct the Barycenter G of two points of different weights m1 and m2.

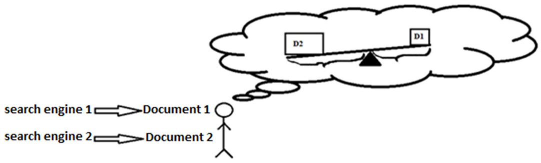


For the balance to be in equilibrium, the momentums m1.GA and m2.GB must be equal. If for example the mass m1 is 4 times greater than the mass m2, it's required to have GA length 4 times smaller than the length GB. This condition results in the following vector equation:

$$\mathrm{m1}.\overrightarrow{\mathrm{GA}}+\mathrm{m2}.\overrightarrow{\mathrm{GB}}=\vec{0}$$

4.4.2 Barycenter of two documents

A user seeking information on the internet, after scanning various search engines, will remain confused between two documents provided by two different search engines. To make his choice, the user has two main factors, namely the documents scores and search engines scores. With these two information, he may consider using physical Barycenter method to make his choice. Indeed, the document score is the attendance rate of this document with respect to the query, i.e., the document weight for the query. The distance between the document and the pivot is represented by the score of the search engine that provided this document. Indeed, a search engine that has a high score is more likely to provide relevant documents and the distance between the object and the pivot should be important to increase the document selection chance.



After putting the two objects in the lever and place the pivot in the correct location, the user will select the winning, ie, the document on the side where there is the balance point (Barycenter Point). In our example, the equilibrium point can be found in the left of the pivot so this is the document D2 that will be chosen.

In the orthonormal (O, $\vec{i}$, $\vec{j}$), we consider the point O as pivot. Both D1 and D2 are placed on each side of the point O such as | OD1 | and | OD2 | respectively represent scores of search engines M1 and M2 (SM1, SM2), the masses represent the documents scores D1 and D2 (SD1, SD2). The equilibrium point G is deducted from the previous vector equality.

$$\mathrm{sd1}.\overrightarrow{\mathrm{GD1}}+\mathrm{sd2}.\overrightarrow{\mathrm{GD2}}=\vec{0}$$

$$\mathrm{sd1}.\overrightarrow{\mathrm{GO}}+\mathrm{sd1}.\overrightarrow{\mathrm{OD1}}+\mathrm{sd2}.\overrightarrow{\mathrm{GO}}+\mathrm{sd2}.\overrightarrow{\mathrm{OD2}}=\vec{0}$$

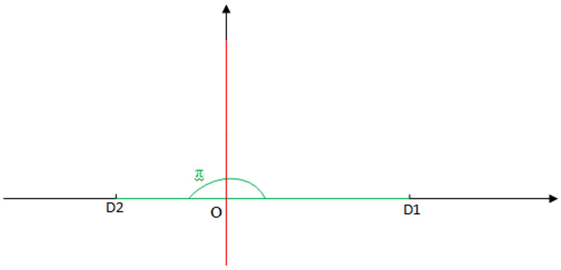
$$(\mathrm{sd1}+\mathrm{sd2}).\overrightarrow{\mathrm{GO}}+\mathrm{sd1}.\overrightarrow{\mathrm{OD1}}+\mathrm{sd2}.\overrightarrow{\mathrm{OD2}}=\vec{0}$$

$$\mathrm{sd1}.\overrightarrow{\mathrm{OD1}}+\mathrm{sd2}.\overrightarrow{\mathrm{OD2}}=(\mathrm{sd1}+\mathrm{sd2}).\overrightarrow{\mathrm{OG}}$$

Point G can be deducted with the following formula:

$$\overrightarrow{\mathrm{OG}}=\frac{\mathrm{sd1}.\overrightarrow{\mathrm{OD1}}+\mathrm{sd2}.\overrightarrow{\mathrm{OD2}}}{(\mathrm{sd1}+\mathrm{sd2})}$$

We can represent the mark by:



If $\overrightarrow{\mathrm{OG}}>0$ then D1 is picked.

If $\overrightarrow{\mathrm{OG}}<0$ then D2 is picked.

Last, we remark that θ1 = 0 (D1 angle) and θ2 = $\frac{2\pi}{2}$ (D2 angle).

4.4.3 Barycenter of three documents

In the orthonormal (O, $\vec{i}$, $\vec{j}$), we consider the point O as pivot. The documents D1, D2 and D3 are represented in polar coordinates by:

- For document D1, θ1 = 0 and |OD1| = Sm1, so xD1=|OD1|.cos(θ1) and yD1=|OD1|.sin(θ1),
- For document D2, θ2 = $\frac{2\pi}{3}$ and |OD2| = Sm2, so xD2=|OD2|.cos(θ2) and yD2=|OD2|.sin(θ2),
- For document D3, θ3 = $\frac{4\pi}{3}$ and |OD3| = Sm3, so xD3=|OD3|.cos(θ3) and yD3=|OD3|.sin(θ3).

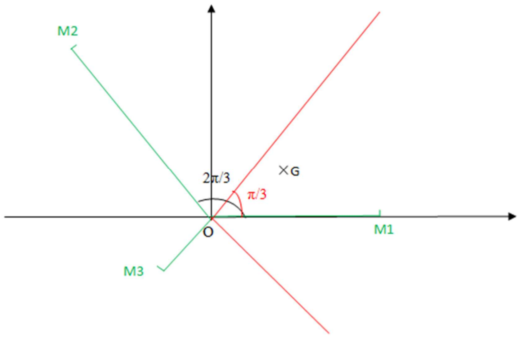
From the previous vector equality, the equilibrium point is deduced by the following formula G:

$$\overrightarrow{OG} = \frac{sd1.\overrightarrow{OD1} + +sd3.\overrightarrow{OD3} + sd2.\overrightarrow{OD2}}{(sd1+sd2+sd3)} = x\,\vec{i} + y\,\vec{j}$$

The polar coordinates of point G are $|OG| = \sqrt{x^2 + y^2}$ and the angle θ of the point G is given by:

$$\begin{cases} \cos(\theta) = \dfrac{x}{|OG|} \\ \sin(\theta) = \dfrac{y}{|OG|} \end{cases}$$

If $-\frac{\pi}{3} < \theta \leq \frac{\pi}{3}$ then D1 is picked.
If $\frac{\pi}{3} < \theta \leq \pi$ then D2 is picked.
If $\pi < \theta \leq \frac{5\pi}{3}$ then D3 is picked.
Example:



In this example the point G is present between $-\frac{\pi}{3}$ and $\frac{\pi}{3}$ so the D1 document will be selected.

4.4.4 Barycenter of many documents

We propose to generalize this method so that the user can make a choice between several documents provided by many search engines.

In the case of multiple search engines, we provide IR2 plan with the orthonormal $(O, \vec{i}, \vec{j})$. Dj document polar coordinates are:

- $\theta i = (j-1)\frac{2\pi}{n}$
- $|ODi| = Smi$
- $xDi = |ODi|.\cos(\theta i)$ et $yD2 = |ODi|.\sin(\theta i)$.

Thus:

$$\sum_{i=1}^{n} Sdi.\overrightarrow{ODi} = \left(\sum_{i=1}^{n} Sdi\right).\overrightarrow{OG}$$

Thereafter:

$$\overrightarrow{OG} = \frac{\sum_{i=1}^{n} Sdi\overrightarrow{ODi}}{\sum_{i=1}^{n} Sdi} = x\,\vec{i} + y\,\vec{j}$$

The G point polar coordinates are $R = \sqrt{x^2 + y^2}$ and the angle θ is given by:

$$\begin{cases} \cos(\theta) = \dfrac{x}{R} \\ \sin(\theta) = \dfrac{y}{R} \end{cases}$$

Finally:

If $\frac{\pi}{n} + (i-2)\frac{2\pi}{n} < \theta i \leq \frac{\pi}{n} + (i-1)\frac{2\pi}{n}$ then Di is picked.

4.4.5 Fusion based on Barycenter

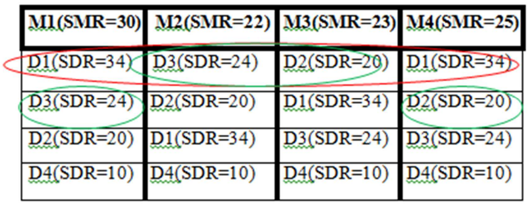
The meta search engine is requested to provide, from a set of ordered documents provided by various search engines list, an ordered list of documents considered most relevant. It is proposed to apply the physical Barycenter method on several results lists to classify documents by relevance.

We will treat an example for applying this method to merge the search engines results and get the documents list considered more relevant.

Example

Considering 4 search engines M1, M2, M3 and M4 which has respectively 30, 22, 23 and 25 as query scores. To apply the Physics Barycenter method, we chose to treat the first 4 results from each search engine (D1, D2, D3, D4), each document Di will have a score SdR(Di).

Table 1 : results of different search engines



M1(SMR=30)	M2(SMR=22)	M3(SMR=23)	M4(SMR=25)
D1(SDR=34)	D3(SDR=24)	D2(SDR=20)	D1(SDR=34)
D3(SDR=24)	D2(SDR=20)	D1(SDR=34)	D2(SDR=20)
D2(SDR=20)	D1(SDR=34)	D3(SDR=24)	D3(SDR=24)
D4(SDR=10)	D4(SDR=10)	D4(SDR=10)	D4(SDR=10)

We apply the Barycenter method by levels. At first we calculate the Barycenter of the first level (D1, D3, D2, D1) which correspond to the documents classified first by the search engines. The nearest document to the Barycenter is considered the leader of this level, in our case the leader of the first level is D1. Thus, it is no longer considered in future calculations.

The second level (D3, D3, D2, D2) contains the 3 remaining documents plus the document which follows document D1 .In our case the leader of the second level is D2. The same principle is applied for the remaining levels.

Finally the meta-search engine classification corresponds to the leaders of each level, in our case the classification is as follows: D1, D2, D3, D4.

5. EVALUATION

To experimentally evaluate the performance of our fusion method results described in this article, we chose the web page collection used in the ninth TREC conference corpus named TREC9. We relied on two measures commonly used in classification, recall and precision.

5.1 Measures used

We use two measures, recall and accuracy, this is the "rate of return", ie the ratio between the number of relevant documents found during a search and the total number of existing relevant documents. The other indicator is the "accuracy rate" which is the ratio between the relevant documents number found during a search and the total documents number found in response to the question. These two concepts are often used because they reflect the user point of view: if precision is low, the user will be dissatisfied because he'll waste time reading information that is not interesting. If the recall is low, the user will not have access to information they wished to have.

5.2 TREC Collection

Since there is currently no standard framework to evaluate a personalized access model to information, we propose an evaluation framework based on "TREC Collections"(Text Retrieval Conference), it is a American conference whose purpose is to allow comparison between the performances of information retrieval systems that exploit large volumes of data, it brings together toolkits and software information retrieval (in full text) designers. It has become a reference and an international standard in the field of information evaluation.

We chose to evaluate our model using the TREC9 collection, it includes 1692096 web pages written in English for a volume of 11,033 MB.

5.3 Learning phase

As a first step, we need to enrich our knowledge base. For this, we launched 10,000 applications to build the knowledge base.

5.4 Experimental results

We measured our approach by 1000 query of several areas, the following figure shows the results for both precision and recall measures. The first tests presented in this figure are very encouraging. The comparison of our approach with existing ones shows that our approach is competitive knowing that our knowledge base is powered over water so the results will be progressively more relevant.

Table 2: Evaluation precision / recall

Nombre de requête	Précision	Rappel
100	0,8563	0,8793
200	0,8571	0,8850
300	0,8537	0,8765
400	0,8543	0,8793
500	0,8596	0,8860
600	0,8775	0,8765
700	0,8793	0,8783
800	0,8680	0,8810
900	0,8851	0,8810
1000	0,8799	0,8820

The relevance evaluation of our meta-search engine is being developed. In this article context, we conducted preliminary experiments to give a rough idea about the quality of our fusion method.

6. CONCLUSION AND PERSPECTIVES

We presented through this paper a method that represents a custom merge using the physical Barycenter method in the results ranking in meta search engine. Thus, we took into account all factors, namely the document score, the search engine score and rank document proposed by the various search engines. We also conducted experiments to evaluate the performance of our meta search engine.

Various improvements can still be proposed, one of our goals is to extract query concepts to treat user query semantically, so we can enrich the query concepts extracted before it is sent to search engines.

REFRENCES:

[1] I.Abdelbaki, Z.Rachik, E.Ben lahmar, E.Labriji, *Int.J.Computer Technology & Applications*,Vol 4 (3), 2013,414-418.

[2] I.Abdelbaki, E.Ben lahmar, E.Labriji, (IJCSIT) *International Journal of Computer Science and Information Technologies*, Vol 4 (2) , 2013, 194 – 198.

[3] R. Mghirbi, K. Arour, Y. Slimani et B. Defude, « Un modèle comportemental d'interclassement de résultats dans un système de recherche d'information P2P », *Actes du XXVIII° congrès INFORSID*, Marseille, mai 2010.

[4] YeeW. G., Frieder O., ≪ On search in peer-to-peer file sharing systems ≫, SAC '05 : *Proceedings of the 2005 ACM symposium on*

Applied computing, ACM, New York, NY, USA, p. 1023-1030, 2005.

[5] Jacques Savoy, Yves Rasolofo, Faïza Abbaci, « Fusion de collections dans les métamoteurs », *JADT 2002 : 6es Journées internationales d'Analyse statistique des Données Textuelles*.

[6] Glover, G.W. Flake, S. Lawrence, W.P. Birmingham, A. Kruger, C.L. Giles, et D.M. Pennock. Improving category specific web search by learning query modifications. Dans *Proceedings of Symposium on Applications and the Internet*, pages 23–31, January 2001. (Cité page 32.).

[7] Beitzel, S. M., E. C. Jensen, A. Chowdury, D. Grossman, O. Frieder et N. Goharian. 2004, On fusion of effective retrieval strategies in the same information retrieval system _, *Journal of the American Society of Information Science & Technology*, vol. 50, no 10, p. 859_868.

[8] Wahlster W. et Kobsa A.,Dialogue-based user models. In *Proceedings of IEEE*, Vol. 74(7), pp. 948-960, 1986.

[9] Challam V., Gauch S., Chandramouli A., « Contextual Search Using Ontology-Based User Profiles», *Proceedings of RIAO 2007, Pittsburgh USA*, 30 may - 1 june, 2007.

[10] Gowan J., A multiple model approach to personalised information access, Master thesis in computer science, Faculty of science, Université de College Dublin, February, 2003. TREC, ≪ Text REtrival Conference ≫, 2008.

[11] Shen D., Chen Z., Yang Q., Zeng H., Zhang B., Lu Y., MaW., «Web-page classification through summarization », In Proceedings of the 27th Annual International ACMSIGIR *Conference on Research and Development in Information Retrieval*, Sheffield, South Yorkshire, UK, p. 242-249, 2004.

[12] Tamine L., Zemirli W., Bahsoun. W., « Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information », *Information - Interaction - Intelligence, Cépaduès Editions*, 2007c.

[13] H. Zargayouna, S. Salotti, et C. Golbreich. Raisonnement par similarité pour l'indexation et la recherche dans des documents multimédia. In *Complément des actes des journées francophones d'Ingénierie des Connaissances* (IC'2001), 2001.

[14] H. Zargayouna et S. Salotti. Mesure de similarité sémantique pour l'indexation de documents semi-structurés. In *12ème Atelier de Raisonnement à Partir de Cas*, 2004.

[15] Sanderson, M. 1994, « Word sense disambiguation and information retrieval", SIGIR 1994, *proceedings of the 17th Annual ACM SIGIR Conference on Research & Development in Information Retrieval*,p.142_151.

[16] Ma Z., Pant G., Sheng, « Interest-based personalized search », *ACM Transactions on Information Systems*, 2007.

[17] Schwartz C. Web Search Engines, *Journal of the American Society for Information Science*. vol. (49):973-982, (1998).

[18] Robertson S. E., Walker S., Hancock-Beaulieu M. M. Large Test Collection Experiments on an Operational, Interactive System: OKAPI at TREC. *Information Processing & Management*, vol. (31):345-360. (1995).