Possibilistic Model for Relevance Feedback in Collaborative Information Retrieval

Fatiha Naouar, Lobna Hlaoua, Mohamed Nazih Omri
Mars Research Unit
Department of Computer Sciences
Faculty of sciences of Monastir
University of Monastir
Monastir, 5000, Tunisia
{fatihanaouar, lobna1511}@yahoo.fr, MohamedNazih.omri@fsm.rnu.tn

ABSTRACT: *Web information is too heterogeneous that users have difficulties to retrieve their needed information: text, image orvideo. In this context, the collaborative work presents one solution proposed to solve this problem. Collaborative retrieval enables the retrieval histories' sharing between users having the same profile across multiple tools such as annotations. We propose in this paper to improve collaborative retrieval performance, considering the annotations as a new source of information describing documents. In our contribution, we propose to apply the relevance feedback to extend the user's query. So we use a possibilistic approach to extract the relevant terms from annotations given in semi-structured documents returned by collaborative retrieval systems.*

Keywords: Collaborative Retrieval System, Possibility Theory, Annotation, Relevance Feedback

Received: 1 March 2012, Revised 15 April 2012, Accepted 21 April 2012

1. Introduction

Facing the vast mass of information found on the web today, a collaborative work's necessary to help the user find his needs. This evolution has improved the performance of retrieval especially for the number of relevant information found and the time put to perform the retrieval. In effect, Working in collaboration allows you to partake the search history as well as formulate a query for collaboration. The collaborative work can be done through multiple tools, in particular considering the annotations which represent relevant information in relation to the document that are to allocate a collection of keywords. In spite of this collaborative framework, the user usually suffers when searching the information to satisfy his needs, which is usually poorly expressed through his query which is composed of simple keywords due to his modest knowledge. In this framework, we suggest to improve the performances of the collaborative retrieval by applying the relevance feedback to enrich the original query. This technique, which consist in extracting terms, starting from documents considered relevant and consider them in a new extended query, was already applied in classic IR [21] and in semi-structured Information retrieval [22] [11] and showed its interest. In our contribution we consider annotations as a new source of information since it allows description of the document by personal users' judgments. The annotation is relatively relevant since it can be made by specialists or not-specialist.

So the relevance feedback using annotations in a collaborative frame brings us back to resolve principally two problems: to known the choice of annotations which can be judged as valid data to consider and the retrieval of the relevant terms which can be re-injected to extend the query.

Several retrieval works were interested in the validation of annotations. We focused this work on the retrieval of the relevant terms used in the annotations to be valid. To do it, we propose a possibilistic model for express the necessity and the possibility of relevance of the terms to be extracted.

We present in the following section a related work on the methods developed for a better collaborative retrieval. We describe our possibilistic model for the retrieval of information in section 3. Then in the section 4 we represent experimentation and results and we conclude at the end.

2. Retated Works

The technological developments to collaborative systems have demonstrated their performance in several areas particularly in the IS [3] [7] [8] [9] [10], in particular, the number of relevant information found and the time taken to perform the retrieval. According to the article published by Lazonder [13], "*two heads look for better than one*". The collaborative retrieval reduces the search time by sharing other results already retrieved. In fact, the collaborative work can be carried out in a Synchronous way through the instantaneous messages or in an Asynchronous way through the electronic mail and the annotations. Collaborative retrieval reduces the time retrieval carried by users having same profile. In addition, il allows to formulate the collaborative queries by the discussion and the queries consultation with the retrieved results.

In this context, the annotations are a popular tool used to share the retrieval results and personal judgments.

According to several works, the annotation can be performed by the content of the document or from an external source to the document [15]. The works annotate by the content of the document (called also indexing) focus generally on the retrieval of the terms. This approach can be base on the classical technique that consists in to attribute a set of keywords (or terms) to every document, or on the semantic technique that attributes a based annotation on concepts (and not simple keywords) and on the relations between them.

We find in the works of Khelif [12] that they take into account the semantic relations between the terms. Njmogue and Al. [18] proposed an approach based on an external source: "*a professional reference*". The principal idea is that, the document indexing depends on the activities of the business and not only on the document terms. This approach uses both a linguistic and statistics analysis of the document and a semantic treatment.

Despite several retrieval systems use the annotations to facilitate the access to exact information; these systems can give results relatively performance since the annotations can be performed by specialists' users or non-specialists. In an even group user one can find experts of the non-experts that can play both of them the role of "*annotato*r". In this context, several retrieval works were performed to find out whether the annotations are judged "*correct*" or not [5].

To ameliorate the performances of the collaborative retrieval we find that the main approaches are based on the history of the researches already made in the group. We find the Spider Collaborative system developed by Chau [6] which allows the user to have access to other researches, for the selection of the best results similar to its needs. Razan [20] suggested a system of support which allows the user to reformulate his query based on the results and "*feedback*" of the collaboration group. The results found are very dependent on the opinions of the group members. The collaborative system SearchTogether proposed by Morris and Horvits [16] bases itself on the safeguard of the websites visited by three people of an even group and of the added in their favorite lists. This system uses the collaboration at various stages in the retrieval process, in the reformulation of queries and in the display of the retrieval results but these notations are limited: the choices are only binary. Other works considered the profiles of the users: we find the works of Naderi and al. [17] which considered that the needs of a user depend not only on his query but also on its profile. They then calculated a similarity of the users' profiles to filter the results already found ameliorate the performances of a collaborative retrieval system. The work of Vivian and Dinet is based on the user's behavior [24]. We note that the problem always returns to a problem of extracting data which several works have been developed. That may be mentioned the work of Omri [19] which is based on the flexible knowledge Extraction Systems and are able to deal with the inherent vagueness and uncertainty of the extraction process.

In their work they used an application that allows a set of collaborators to work on the same thematic [23]: it allows the notation of the visited pages, the visualization of the notes already attributed by the group of collaborators, the display of the list of classified pages in a given retrieval thematic and the restitution of the notes in system retrieval pages.

This application showed a possibility of optimization of the information retrieval but its evaluation was carried out with restrict group of people (18 students). Armin suggested is calculating the global relevance from the previous query [2]. It's a question of calculating the similarities of a new query and each of the existing query and relevant documents stemming from researches corresponding to most of the similar query. Other studies have tried to enrich the original query by adding the terms selected from a collaborative website [14] the n most relevant tags returned by the system. The results showed a slight improvement.

3. A Possbilistic Model for Information Retrieval

3.1 Motivation

Our objective is to enrich the initial query composed of simple keywords of the user, by relevance feedback to be able to find the appropriate information. In view of the importance of the annotations in collaborative Retrieval, we have considered them new source of information which can really describe the document. So, our approach consists in extracting the pertinent terms from annotations to express better the need of the user. But the relevance of a term is not certain and we speak then of a degree of relevance that expresses a user's preference.

By examining annotations we can distinguish two types of appearance of terms: some appear in titles, other are and found only in the body of documents. This is why we thought to discriminate the two types of appearance by considering a dual measure' found in possibility theory: The necessity is intuitively connected to the terms appearing in the titles and the possibility connected to the terms appearing in the body of the document. Our model is then based on a possibilistic network allowing introducing the different relations of dependency. This approach will be detailed in the next section.

3.2 Architecture of the model

We suggest a model based on a possibilistic network (figure 1). The nodes represent the documents D composed of an annotation A, a title T and a body of the text Tx. These elements contain terms which can belong to the query R. The arcs represent the relations of dependency. X_i nodes can represent elements "*annotation*" , "*titre*" or "*texte*" Each node X_i is a binary random variable taking values in the set dom $(X_i) = \{x_i, \neg x_i\}$. The instantiation $X_i = x_i$ (resp. $X_i = \neg x_i$) means that the element x_i is relevant (resp. irrelevant) to the document D. Each node Mi represents a binary random variable taking values in the set dom $(M_i) = \{m_i, \neg m_i\}$. The instantiation $M_i = m_i$ (resp. $M_i = \neg m_i$) means that the term is representative M_i (resp. not representative) of the parent node connected to it. Every variable Ai, Ti and Tx depends directly on its parent node is the root node D in the possibilistic network. So, every variable at the M_i, M_i $M = \{M_1, M_2, \ldots, M_n\}$ depends only on its parent node can be a variable or text annotation or title. They consider that M(X) represents the set of terms constituting X where X can be a title, an annotation, a query or the rest of the text.

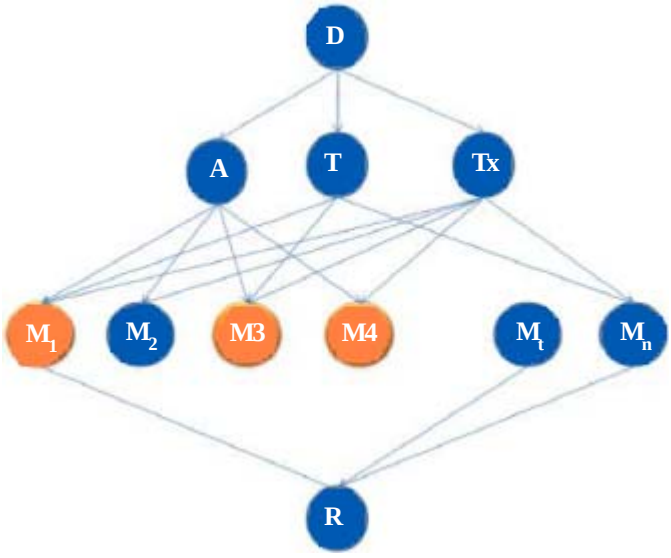

Figure 1. Model based possibilistic network

3.3 Calculation of relevance depending on the possibilistic model

We base the necessity and the possibility of a term by the following two definitions:

Definition 1: A term of the annotation is considered possibly to extract if it appears frequently in the body of the text of the documents.

Definition 2: A term of the annotation is considered necessarily to extract if it appears frequently in the titles of the documents. Relevance is defined by two dimensions according to the possibility theory [25]: the necessity and the possibility. The necessity translates that a term from the annotation belonging to a title is necessarily relevant. While a term belonging to the annotation bellowing to the text is possibly relevant to the reformulation of the query.

To evaluate the degree of possibility of a term knowing that it appears in a body of text $\Pi(M_j / Tx)$ and the necessity degree $N(M_j / T)$, of a term knowing that it appears in a title, we were inspired by the model developed by Brini [4]:

$$\Pi(M_j / Tx) = (\Pi(Tx \wedge M_j)) / \Pi(Tx) \quad (1)$$

$$N(M_j / T) = 1 - \Pi(\overline{M}_j | T) \quad (2)$$

$$\text{Avec } \Pi(\overline{M}_j | T) = (\Pi(T \wedge \overline{M}_j)) / \Pi(T) \quad (3)$$

The relevance of a term is calculated by varying the necessity and the possibility with two factors α and β as:

$$Rel(t_1) = \alpha * N(t_1) + \beta * P(t_1) \quad (4)$$

With α and $\beta \in [0, 1]$ and $\Sigma \alpha + \beta = 1$

3.4 Retrieval of the terms

Our objective is to ameliorate the performances of the collaborative retrieval systems by using the valid annotations of the documents returned by a retrieval system. Our contribution consists mainly in selecting the correct annotations of the relevant documents resulting from a user query to extract the relevant terms in our possibilistic model, reformulate and compare. To do this we have to pass by the indexation of documents to process. These documents can be in the type of text or picture.

To give a sense of representativeness of a term, of a valid annotation for a given relevant document, we used a combination of factors tf $*$ ief. The frequencies of the terms of a given document are interesting to measure to what extent an element is exhaustive while the inverse frequency allows measuring to what extent a term is specific of the collection.

For every term of the annotation, we calculated their occurrence numbers in elements e_i title, the size of elements e_i and its appearance in the number of terms.

$$tf(t_1, e_i) = \frac{\sum i = 1...n \, Occ(t_1, e_i)}{taille(e_i)} \quad (5)$$

With t_1 represents a term of the annotation and e_i represents an element : e_i = {title, body}.

The value of ief of the term t_1 of all the elements E_i of all collection is performed by the following expression:

$$ief(t_1, E_i) = log \frac{|E_i|}{|\{e_i \in E_i : t_1 \in e_i\}|} \quad (6)$$

With $|E_i|$ The cardinality of the total number of the elements in the collection and $|\{e_i \in E_i : t_1 \in e_i\}|$ is the cardinality of the number of the element or the term t_1 appears (that is to say $tf(t_1, e_i) \neq 0$). If the term is not in the collection and for not to divide by zero thus we change the formula and becomes:

$$ief(t_1, E_i) = log \frac{|E_i|}{1 + \{e_i \in E_i : t_1 \in e_i\}} \quad (7)$$

The calculation of $tf * ief$ is made according to the following formula:

$$tf * ief = tf(t_1, e_i) * ief(t_1, E_i) \quad (8)$$

3.5 Calculation of the necessity and the possibility degrees

We calculated the necessity and the possibility degree of a term of an annotation with two concepts: the notion of frequency *tf* and the notion of $tf * ief$. For a term of annotation t_1, its necessity degree is determined according to its appearance in the title, e_i = **{*tittle*}**. The necessity degree is calculated in two ways:

$$N(t_1) = \frac{\sum i = 1...n \; tf(t_1, e_i)}{\sum i = 1...n \; tf(e_i)} \quad (9)$$

$$N(t_1) = \frac{\sum i = 1...n \; tf(t_1, e_i) * ief(t_1, e_i, E_i)}{\sum i = 1...n \; tf(e_i)} \quad (10)$$

The possibility of a term t_1 of an annotation is determined by its appearance in the body of the text, e_i = {body}. The Possibility is calculated in two ways:

$$P(t_1) = \frac{\sum i = 1...n \; tf(t_1, e_i)}{\sum i = 1...n \; tf(e_i)} \quad (11)$$

$$P(t_1) = \frac{\sum i = 1...n \; tf(t_1, e_i) * ief(t_1, e_i, E_i)}{\sum i = 1...n \; tf(e_i)} \quad (12)$$

4. Experimental Study and Results

Our objective in this section is to evaluate the impact of the relevance feedback based on annotations.

The literature is lacking of evaluation system in collaborative retrieval, we considered the following tools of evaluation:

- A collaboratif retrieval system: We chose the "*YouTube*" which is a popular and social system and which allows to give a collaboratif service of annotation of means MM.
- A collection of document: It will be constructed by the documents returned by the system itself and that will be saved for retrieval of different part already described in previous sections.
- Queries: We have proposed four queries on different domains.
- The relevance judgments are made by ourselves.
- Performances are measured by the precision for 5, 10 and 20 top documents by using the residual relevance feedback.
- The results acquired for initial queries are summed up in the following table:

	5 docs	**10 docs**	**20 docs**
Q1= « unix server »	0.2	0.2	0.2
Q2= «esthetic surgery »	0.4	0.3	0.25
Q3= « brain cancer »	0.0	0.0	0.25
Q4= « history alien »	0.2	0.4	0.4
Average	0.2	0.22	0.27

Table 1. Precision of Initial Queries

For the reformulation of queries we performed a manual retrieval of the valid annotations of the first two relevant documents returned by the system.

Then we calculated, the degree of necessity by using $tf * ief$ value which is the frequency of elements inverse (10), as well as the degree of possibility of each term of annotations.

To reformulate a query, we added the term with the highest score without taking into account the terms which already appear in initial query.

	Queries	5 docs	10 docs	20 docs
Necessity	Q1	0.2	0.2	0.2
	Q2	0.6	0.4	0.3
	Q3	0.6	0.4	0.45
	Q4	0.4	0.4	0.45
Average		0.45	0.35	0.35

Table 2. Precision of the queries reformulated by the term most Necessary

According to the Table 2, we see that the precision is improved as compared to the original query in particular in the top 5 documents returned. It attains an average of 0.45 against 0.2 by using initial query. What confirms the interest of annotations and terms that appear in the tags "*title*".

	Queries	5 docs	10 docs	20 docs
Possibility	Q1	0.4	0.3	0.35
	Q2	0.6	0.4	0.25
	Q3	0.6	**0.7**	0.5
	Q4	0.6	0.5	0.5
Average		0.55	0.47	0.4

Table 3. Precision of the reformulated queries by the term as possible

According to the Table 3, we see that the precision is improved as compared to the original query in particular in the top 5 documents also returned. It attains an average of 0.55 against 0.2 by using initial query. What confirms the interest of annotations and terms that appear in the tags "*body*".

Generally, the improvement of precision by using the possibility is lightly more important than the case using the necessity. This can be explained by the fact that the body of the text from which we extracted the relevant terms (12) is richer in information than the title.

We also calculated the precision rate of aggregation with $\alpha = 0.5$ and $\beta = 0.5$ for the reformulation of the four queries Q1, Q2, Q3 and Q4 and we calculated the average relevance's. In table 4 we represent the precision rate by adding a single term.

Calculion	5 docs	10 docs	20 docs
Relevance	0.55	0.47	0.4
Possibility	0.55	0.47	0.4
Necessity	0.45	0.35	0.35

Table 4. Average precision of reformulated by adding a single term

Noted that there is no real improvement since we tested with the addition of a single term and then the result corresponds either to adding one term depending on the necessity or possibility. To better exploit the aggregation of two measures, we will test the addition of one more term.

To better to see the behavior of our algorithm, we tried to reformulate initial query by adding two terms by exploiting the aggregation between the necessity, the possibility and the both. The calculation of precision rate of the reformulation of the four queries is represented in Table 5.

Noted that there is no real improvement since we tested with the addition of a single term and then the result corresponds either to adding one term depending on the necessity or possibility. To better exploit the aggregation of two measures, we will test the addition of one more term.

To better to see the behavior of our algorithm, we tried to reformulate initial query by adding two terms by exploiting the aggregation between the necessity, the possibility and the both. The calculation of precision rate of the reformulation of the four queries is represented in Table 5.

Calculion	5 docs	10 docs	20 docs
Relevance	0.6	0.6	0.5
Possibility	0.6	0.6	0.5
Necessity	0.45	0.35	0.35

Table 5. Average precision of reformulated by adding two terms

We noted that there is an interesting improvement for the first 5 documents returned with an average equal to 0.6 compared with the precision of the initial queries, which is 0.2.

We note from the last two tables (table 4 and table5) there is a better improvement on the first 5 documents returned by the system and by exploiting the aggregation of two measures: the necessity and the possibility.

The comparison of the rates of precision of the different measures is represented in Figure 2.

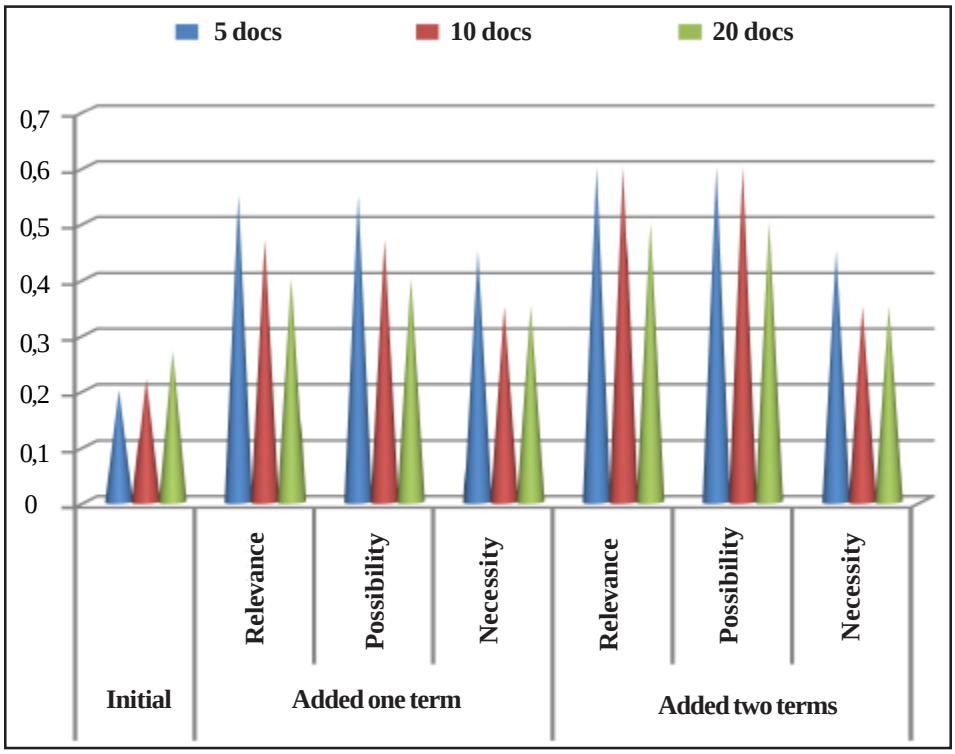

Figure 2. Comparison of precision rate

Note that the reformulation of two most plausible and the term most necessary increase the precision rate of documents returned by the system compared with the reformulation by a single term.

To evaluate our system, we calculated rate for different queries:

$$\textit{Improvement rate} = \frac{\textit{New precision - old precision}}{\textit{old precision}} \quad (13)$$

It is represented in Figure 3.

There is a better improvement on the first 5 document returned by the system by the addition of two terms that can reach 20%.

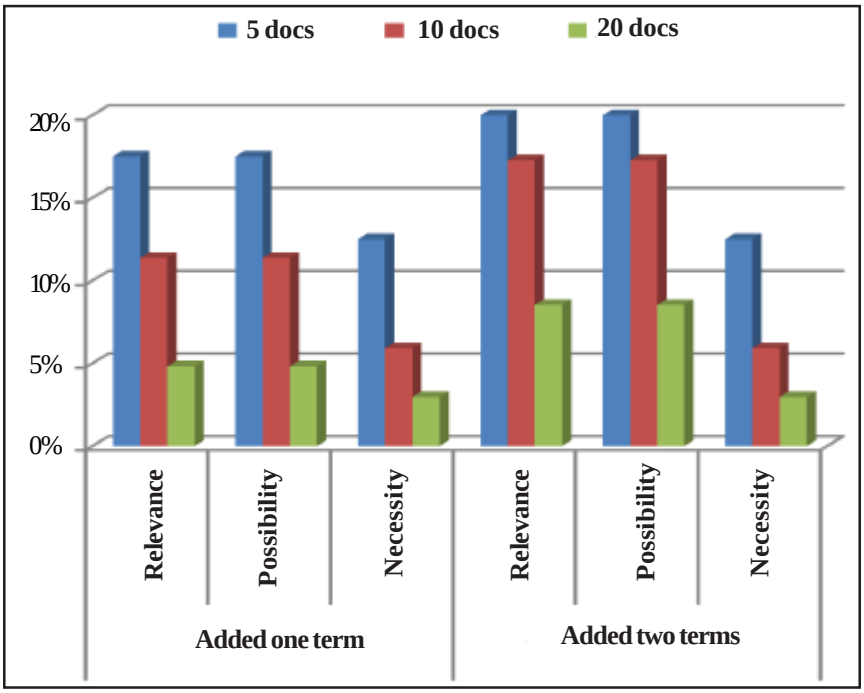

Figure 3. Rate of improvement

5. Conclusion and Future Works

We proposed a new approach to possibilistic relevance feedback in collaborative information retrieval. We considered the "*valid"* annotation as a source of information. We evaluated our approach by calculating the precision rate and the improvement rate for 5, 10 and 20 top documents. The results are encouraging since we have found a better improvement with the aggregation of terms especially for the first 5 documents, which address the need for the user. We tested an aggregation between possibility and necessity. We propose in our future work to expand our test database. We also propose to create a model validation.

References

[1] Abrouk. L. (2006). Annotation de documents par le context de citation basée sur une ontology, *Thèse de doctorat en. informatique, Montpellier.*

[2] Armin, H., Stefan, K., Markus, J., Andreas, D. (2003). Towards collaborative information retrieval: Three approaches. *In Ingrid Renz Jurgen Frank, Gholamreza Nakhaeizadeh, editor, Text Mining – Theoretical aspects and Applications*. Physica-Verlag.

[3] Bharat, K. (2000). SearchPad: explicit capture of search context to support web search, *WWW9, In: Proceedings of the Ninth International World Wide Web Conference*, Amsterdam The Netherlands, May, p. 15—19.

[4] Brini, A., Boughanem et, M., Dubois, D. (2006). Réseaux possibilistes pour un modèle de recherche d'information. *In: Proceedings of CORIA'2006*. p.143-154.

[5] Cabanac, G. (2008). Annotation collective dans le contexte RI : définition d'une plate-forme pour expérimenter la validation sociale, *CORIA 2008 : Conférence en Recherche d'Information et Applications*, p.385-392.

[6] Chau, M., Zeng, D., Chen, H., Huang, M., Hendriawan, D. (2003). Design and Evaluation of a Multi-agent Collaborative Web Mining System. *Decision Support Systems (DSS)*. 35 (1) 167-183.

[7] Cockburn, A. et, McKenzie, B. (2001). What do web users do ? An empirical analysis of web use, *International Journal of Human-Computer Studies*, 54, 903-922.

[8] Diamadis, E. T. et, Polyzos, G. C. (2004). Efficient cooperative searching on the Web: system design and evaluation, *International Journal of Human-Computer Studies*, 61, 699-724.

[9] Dinet, J. (2007). *Deux têtes cherchent mieux qu'une ?*, Medialog, 63.
[10] Dumais, S., Cutrell, E. Chen, H. (2001). Optimizing search by showing results in context, *In*: CHI'01, Proceedings of the ACM Conference on Human-Computer Interaction, Seattle, USA, March 31 – April 5, New-York, ACM Press, p. 277-284.
[11] Hlaoua, L. (2006). Reformulation de requêtes par structure en RI dans les documents XML, CORIA, Lyon.
[12] Khelif, K., Dieng-Kuntz, R., Barbry, P. (2007). An ontologybased approach to support text mining and information retrieval in the biological domain, *Special Issue on Ontologies and their Applications of the Journal of Universal Computer Science (JUCS)*, 13 (12) 1881-1907.
[13] Lazonder, A. (2005). Do two heads search better than one? Effects of student collaboration on web search behaviour and search outcomes, *British Journal of Educational Technology*.
[14] Lioma, C., Moens, M. F., Azzopardi, L. (2008). Collaborative annotation for pseudo relevance feedback , *ECIR workshop on exploiting Semantic Annotation in Information Retrieval (ESAIR 2008)*.
[15] Mokhtari, N. et, Dieng-Kuntz, R. (2008). Extraction et exploitation des annotations contextuelles, *Extraction et gestion des connaissances (EGC'2008)*.
[16] Morris, M. R., Horvitz, E. (2007). SearchTogether : An interface for collaborative Web Search, UIST '07: *In: Proceedings of the 20th annual ACM symposium on User interface software and technology*.
[17] Naderi, H., Rumpler, B., Pinon, J. M. (2007). An Efficient Collaborative Information Retrieval System by Incorporating the User Profile, *In*: Adaptive Multimedia Retrieval: User, Context, and Feedback Lecture Notes in Computer Science.
[18] Njmogue, W., Fontaine, D. et, Fontaine, P. (2004). Identification des thèmes d'un document relativement à un référentiel métier, *In Proceedings of MAJECSTIC'04*, Calais, France.
[19] Omri, M.-N. (2004). Pertinent Knowledge Extraction from a Semantic Network: Application of Fuzzy Sets Theory. International Journal on Artificial Intelligence Tools (IJAIT). 13 (3) 705-719.
[20] Razan, T. (2004). Soutien Personnalisé pour la Recherche d'Information Collaborative. *2ème Congrès MAJECSTIC 2004. Manifestation de JEunes Chercheurs Sciences et Technologies de l'Information et de la Communication*, Calais, France.
[21] Rocchio, J. (1971). Relevance feedback in information retrieval, *The SMART retrieval system-experiments in automatic document processing*, Prentice Hall Inc, p. 313-323.
[22] Schenkel, R., Thbobald, M. (2005). Relevance Feedback for Structural Query Expansion, *INEX 2005 Workshop Pre-Proceedings, Germany*, p. 260-272.
[23] Vivian, R., Dinet. J. (2008a). La recherche collaborative d'information ; vers un système centre utilisateur. *Les cahiers du numérique*.
[24] Vivian, R., Dinet. J. (2008b). Présentation d'un système centré utilisateur d'aide à la recherche collaborative d'information : Présentation des premiers développements d'un outil. *Self 2008*, Ajaccio.
[25] Zadeh, L.A. (1978). Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets and Systems*, 1 (1) 3-28.