RELEVANCE FEEDBACK: INTRODUCTION OF PARTIAL ASSESSMENTS FOR QUERY EXPANSION

Asma H. BRINI
IRIT, Université Paul SABATIER
118, Route de Narbonne
31062 Toulouse CEDEX 04
brini@irit.fr

Mohand BOUGHANEM
IRIT, Université Paul SABATIER
118, Route de Narbonne
31062 Toulouse CEDEX 04
bougha@irit.fr

Abstract

This paper deals with flexible information retrieval. A method of partial relevance feedback is presented. Users give partial preference to document retrieved by the system. Experiments carried out on the TREC10 test collection showed the effectiveness of the method.

Keywords

Information Retrieval, Relevance feedback, Fuzzy Logic, Partial assessments.

1 INTRODUCTION

The objective of Information Retrieval (IR) is to retrieve relevant documents in response to a user need [16] [19]. This need is expressed as a query through keywords or terms in natural language. However, it is often difficult for users to express their needs since they have a vague or imprecise idea of what they want. It is usually easier for them to assess documents retrieved by the initial query. Relevance feedback methods have been proposed for this purpose for different IR models (probabilistic, VSM, connectionist) [14], [10], [15], [5]. Most of these methods consider the relevance concept as binary. The user marks the documents she/he considers as relevant (the others are considered as nonrelevant). Terms extracted from relevant documents are added to the new query. However, the relevance cannot be considered as binary, it is rather dynamic and partial [13], [17], [18], [11], [12].

Our work focuses on relevance feedback (RF) and particularly on partial judgements. We consider that a binary judgement in RF leads to generating a wrong query. Indeed a document which seems partially relevant to the user can be assigned to either the relevant or nonrelevant set of documents. This assignation may bias the constructed final query. For example, a document which seems partially relevant is assumed to be relevant as the user's choice of judgement is limited. As a consequence of this assignation, terms of this document may appear in the final query. The problem is that these terms may actually appear in the non-interesting part of the document. Our goal is to avoid such a mistake and to model and interpret the partial judgements of users through fuzzy logic. We choose fuzzy logic as it well known that it gives a framework that can handle uncertainty and vagueness. Uncertainty and vagueness are intrinsic to the whole process of information retrieval.
This is to be discussed along with motivation and methodology in section 2. In section 3, we give an overview of the experiments we carried out on the TREC collection and conclude in section 4.

2 PARTIAL ASSESSMENTS IN RELEVANCE FEEDBACK

In IR, fuzzy set methods [22] provide a framework for modelling flexibility and handling vagueness and uncertainty which are intrinsic to information. The flexibility is expressed through different processes of IRS. It has been applied to flexible queries and document representation and to a flexible evaluation when matching queries with documents [1], [2], [3], [9]. Systems do not only retrieve documents of which the representation corresponds exactly to the query but also documents that match the query partially. We extend this idea by giving the user the possibility to assess documents that are partially relevant. Instead of giving judgement on document parts such as the title, section etc, the user judges the whole document, giving partial relevance. Assessments are done using linguistic quantifiers. We first translate these judgements using the fuzzy approach and then build a new query that is submitted to the system.

2.1 PARTIAL ASSESSMENTS

Usually, in Classical Relevance Feedback (CRF) method users are asked to judge documents in two degrees of relevance (relevant/nonrelevant). A new query is built by adding terms extracted from relevant documents with weights higher than in the nonrelevant ones. For example, in the VSM [14], the new query is obtained as follows:

$$Q' = \alpha.Q + \frac{\beta}{n_p}\sum_{n_p} D_p - \frac{\gamma}{n_{np}}\sum_{n_{np}} D_{np}$$

where:
Q' is the new query,
Q the initial query,
n_p (respectively n_{np}): the number of relevant (respectively nonrelevant) documents,
D_p (respectively D_{np}): the relevant (respectively nonrelevant) document representation,
α,β,γ: constant.

We propose a new algorithm for query expansion based on partial judgements of retrieved documents. We introduce three levels of judgements: 'Very Relevant' (VR), 'Relevant' (R) and 'Not Relevant' (NR). These linguistic levels are then translated in a mathematical way. We define three fuzzy sets: VR, R, NR representing the relevance levels.

$$\mu_i : J \times D \rightarrow [0,1],\ J = \{VR, R, NR\},\ i \in \{VR, R, NR\}$$

$$VR = \sum_{d\in D} \mu_{VR}(d)/d\ ,$$

$$R = \sum_{d\in D} \mu_R(d)/d$$

and,

$$NR = \sum_{d\in D} \mu_{NR}(d)/d$$

D: documents of the collection,
d: document representation.
The degree of membership of a document to a relevance level is binary: since we totally trust the user's assessment. Then:

$$\mu_{VR}(d) = \begin{Bmatrix} 1 & if\ d \in VR \\ 0 & otherwise \end{Bmatrix},$$

$$\mu_R(d) = \begin{Bmatrix} 1 & if\ d \in R \\ 0 & otherwise \end{Bmatrix},$$

and

$$\mu_{NR}(d) = \begin{Bmatrix} 1 & if\ d \in NR \\ 0 & otherwise \end{Bmatrix}.$$

This first step is to formalise the document judgements; the second is to transfer them onto the terms occurring in these documents. The objective is to build the new query with the best couples (term, weight).

2.2 NEW QUERY BUILDING

The new query is built by considering the terms extracted from the judged document and the judgements of these documents

We define the document representation as a membership function (a term occurs in a document at a certain degree of membership): $\mu_d : D \times T \rightarrow [0,1]$, $d = \sum_{t\in T} \mu_d(t)/t$.

Once the user gives her/his judgement, we extend the document representation to a judgement-based document representation. Therefore, we define a binary relation which indicates the importance of a term *t* in a document *d* of judgement *j*:

$$d = \sum_{(t,j)\in T\times J} \mu_{di}(t,j)/(t,j),$$

where:
J={VR, R, NR},
T: collection terms,
d: document representation,
t: represents a term,
T: total collection terms.

$$\mu_{di}(t, j) = \alpha * \mu_d(t),$$

α depends on the level of relevance decided by the user and by the membership function of a term to different levels of relevance. It is a constant number obtained in an experimental way.

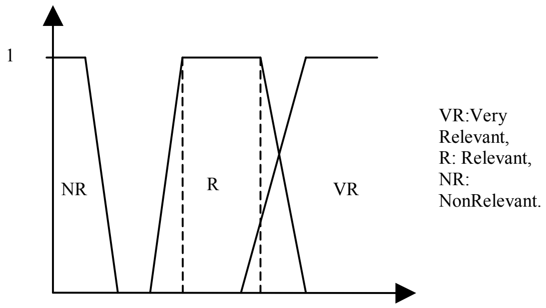

Figure 1 Membership functions of the three relevance levels

The graph above illustrates the relevance levels.
We assume that VR document means that most of its terms are VR, these terms are called VR terms. A document is R if at least between 60% and 90% of its most representative terms are R and NR document is a document in which not more than 10% of its terms are relevant [6]. This assumption gives the value of α.

In order to measure the $\mu_d(t)$ we extended [7] the term weight formula used in the Mercure IRS [4]. This formula is as follows:

$$d_{ij} = \frac{tf_{ij} * \left(h_1 + h_2 * \log\left(\frac{N}{n_i} \right) \right)}{h_3 + h_4 * \frac{dl_j}{\Delta_d} + h_5 * tf_{ij}}$$

Where:
$\mu_d(t)$ =dij
d_{ij} : link weight between term t_i and document j ,
tf_{ij} : frequency of occurrence of term i in document j,
n_i : number of document containing the term t_i ,
dl_j : width of document d_j ,
N: number of documents of the collection,
h_1, h_2, h_3, h_4: constant parameters

Knowing this, a given term may belong to VR, R or NR documents. The final weight of this term is aggregated either by OWA operators [21] or the Maximum/Minimum pair. We choose the Maximum/Minimum pair aggregation operation. We define two classes:

- the relevance levels VR and R define the class of positive terms,
- the class of negative terms is defined by the relevance NR.

The final weight of a term W_t is calculated as following:

$W_t = Max(\mu_{di}(t,j))$, if $t \in$ VR$\vee$ R
$W_t = Min(\mu_{di}(t,j))$, if $t \in$ Positive $\wedge$ NR.

At this stage, a list of all terms occurring in all the judged documents is produced. Usually only X=20 top terms are selected as a final query. We choose to process only one feedback iteration.

3 EXPERIMENTS AND RESULTS

The goal of the experiments is to evaluate the effectiveness of the partial judgement in query expansion. These experiments are carried out on TREC10 test collection provided by TREC programme [20]. This collection contains documents (10 Go of texts), queries (50), and relevance judgments (which documents are relevant to the query). TREC 10 collection provides three relevance levels expressed through numbers (0,1,2).
In order to evaluate the effectiveness of the proposed method called Fuzzy Relevance Feedback (FRF), we compare this method to the Classical (binary) Relevance Feedback (CRF) presented in section 2.1.
The experiment is done as follows:

For each relevance feedback methods

- For each query
  - o submit the query to the IRS (we used our local IRS system called Mercure [4])
  - o the system returns 1000 documents
  - o the top 12 documents are judged according to the considered RF method
  - o a new query is then built according to the considered method
  - o this query is submitted to the IRS which returns a ranked list of 1000 documents.

Thus, at the end of the experiment two lists of 1000 documents are selected for each query. One list is the result of FRF and the other is the list of

CRF. These lists are then evaluated using the standard measure precision developed in IR domain and mostly used in TREC programme [20]. Eight precision measures noted P5, P10, P15, P20, P30, P100, P1000 and AvgPr were used in our experiments. Pn (n=5, 10, …,1000) is the precision at n documents, it is measured as follows *Pn= r/n*. Where, r: Number of relevant document. n: Number of retrieved documents which are kept. *AvgPr* is the average precision at the two related documents.

Notice that the precision values listed in the different tables and figures are the average precision computed on the 50 used queries.

Figure 2 shows the precision curves, at different points P_n.

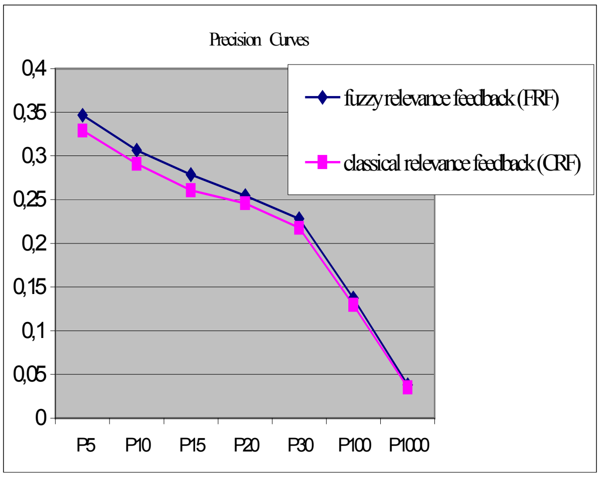

Figure 2 Precision curves for the two methods

The proposed method improves the classical method with two main important results:

i) the retrieved list is always better P_i
ii) the best documents are in the top of the retrieved-document list.

Table 1 shows the different values of Figure 2. The fourth line indicates the improvement measured by:

$$RV_i = ((Pi_{FRF} - Pi_{CRF}) *100),$$

Table 1 Precision values

	P5	P10	P15	P20	P30	P100	P1000	Avg Pr
CRF	0.3289	0.2911	0.2607	0.2456	0.2178	0.1296	0.0352	0.1087
FRF	0.3467	0.3067	0.2785	0.2544	0.2281	0.1373	0.0380	0.1182
V.R *	5.4119	5.3589	6.8277	3.5830	4.7291	5.9413	7.9545	8.7396

*V R:. Variation rate (%) ,
AvgPr: Average Precision.

Where:
RV_i is the variation at precision point i ,
Pi_{FRF} , Pi_{CRF} are the precision at point i of FRF and CRF respectively.

The second experiment compares the FRF to the initial search, i.e. the search performed by the initial query. The goal is to measure how well the partial judgments are benefit comparing to "no feedback". The evaluation method that has been used is called the residual ranking evaluation [8]. In this method the documents that have been seen (judged) are dropped from the both lists (feedback and no feedback) of retrieved documents. Precision measures are then computed. This method tells us how much we gained by doing the feedback comparing to the initial search. Figure 3 compares the FRF, named feedback in the figure, and the initial search.

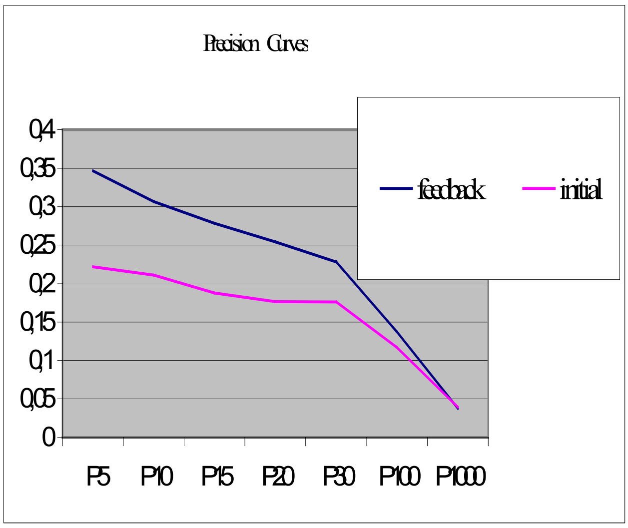

Figure 3. Precision curves for the two methods

As expected, the feedback curve is definitely higher for the points 5, 10, 15, 20, 30.

4 CONCLUSION

The work presented in this paper introduces partial relevance in the process of relevance feedback. A users' partial preference is close to human reasoning. When users have several levels, they can easily make choices to assign documents to relevance levels which best match their idea and perceptions of the documents. Elements of fuzzy logic give an adequate framework to handle these partial judgements. Experiments undertaken on TREC collection show the effectiveness of the approach. A false judgement on a document entails generating a false query and therefore decreases the system's performance. The next work will give users the possibility to judge documents relatively, so as to express that documents d1, d2 are relevant but d2 is better (more relevant) than d1.

References

[1]. P. Bosc, H. Prade (1996). *An Introduction to Fuzzy Set and Possibility Theory-Based Treatment of Soft Queries and Uncertain or Imprecise Databases*. Uncertainty Management in Information Systems, 1996, pp 285-324.

[2] G. Bordogna , G. Pasi (1997). *Application of OWA Operators to Soften Information Retrieval Systems*, The Ordered Weighted Averaging Operators Theory and Applications R.R. Yager, J. Kacprzyk, Kluwer Academic Publishers, 1997, pp 275-292.

[3] G. Bordogna , G. Pasi (2000). *Modeling Vagueness in Information Retrieval*. ESSIR 2000. LNCS 1980, pp 207-241. Springer-Verlag Berlin Heidelberg 2000.

[4] M.Boughanem (1992). *Les Systèmes de Recherche d'Informations : d'un Modèle Classique à un Modèle Connexionniste,* Doctoral Thesis, Toulouse (France), December 1992.

[5] M. Boughanem, Linda Tamine (2000). *Query Optimization Using An Improved Genetic Algorithm*. CIKM 2000: pp 368-373.

[6] A. H.Brini (2002). *Introduction de la Gradulaite dans le Jugement Utilisateur*. DEA Report. Toulouse (France), June 2002.

[7]: D.A. Buell, D.H. Kraft (1987). *TIRS : A Topological Information Retrieval System Satisfying the Requirements of the Waller-Kraft whish list*. Information Processing and Management, 1987, pp 171-180.

[8] Y.K. Chang, C. Cirillo and J. Razon. (1971). *Evaluation of feedback retrieval using modified freezing, residual collection and test and control groups*, chapter 17, pp. 355-370. Prentice-Hall Inc., 1971. In The Smart Retrieval System].

[9] D. Dubois, H. Prade, F. Sèdes. (2001). *Fuzzy Logic Techniques in Multimedia Database Querying: A Preliminary Investigation of the Potentials.* TKDE 13(3), 2001, pp 383-392.

[10] D. Harman. *(1992). Relevance Feedback Revisited, Conference on Research and Development in Information Retrieval* (SIGIR). 1992, pp1-10.

[11] K. J. Järvelin, J.Kekäläinen. *(2000). IR evaluation methods for retrieving highly relevant documents*. In: Belkin, N. & Ingwersen, P. & Leong, M-K. (Eds.), Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR 2000).

[12] K.L. Maglaughlin, D.H. Sonnenwald. (2002). *User Perspectives on Relevance Criteria: A Comparison among Relevant, Partially Relevant, and Not-Relevant Judgements*, JASIST, 53(5): 2002, pp 327-342.

[13]. S. Mizzaro. (1997). *Relevance: the Whole History*, JASIS, 48(9): 1997, pp 810-832.

[14] J.J.Rocchio. (1971). *Relevance Feedback in Information Retrieval*, in The Smart System Experiments in Automatic Document Processing in Automatic Document Processing, ed Prentice-Gall. 1971, pp 313-323.

[15] S.E. Robertson. (1994). Computer retrieval as seen through the pages of Journal of Documentation. In: B.C. Vickery (ed.), *Fifty years of information progress*. Aslib, 1994, pp 119-146.

[16] G. Salton. (1971). *The SMART Retrieval Sytem.* Englewood Cliffs, NJ, Prentice Hall 1971.

[17] A. Spink,H. Greisdorf, J. Bateman. (1998). *From Highly Relevant to Not Relevant: Examing Different Regions of Relevance*, Information Processing and Management, Vol 34, No 5, 1998, pp 599-621.

[18] R. Tang, P. Solomon. (1998). *Toward an Understanding of the Dynamics of Relevance Judgement*: An Analysis of One Person's Search Behavior, , Information Processing and Management, Vol 34, No 2/3, 1998, pp 237-256.

[19] C.J. van Risbergen. (1979). *Information Retrieval*. 2^{nd} Edition. Butterworths. 1979.
[20] E. Voorhees. (2001). *Overview of the TREC 2001.Proceedings of the 10th Text REtrieval Conference TREC-2001,pp 1-15,* November 2001, pp 19-20.
[21] R.R.Yager. (1988). *On Ordered Weighted Averaging Aggregation Operators in Multi-Criteria Decision Making*. IEEE Transactions on Systems, Man, and Cybernetic*s*, 18: 1988, pp183-190.
[22] L. A. Zadeh. (1965). *Fuzzy Sets*, Information Control, 1965, vol. 1 pp 338-353.