International Journal of Fuzzy Logic and Intelligent Systems, vol. 2, no. 1, March 2002 pp. 9-14

A Study on Improving the Effectiveness of Information Retrieval Through P-norm, RF, LCAF

Young-cheon kim and Sung-joo Lee

Department of Computer Science, Chosun University

Abstract

Boolean retrieval is simple and elegant. However, since there is no provision for term weighting, no ranking of the answer set is generated. As a result, the size of the output might be too large or too small.

Relevance feedback is the most popular query reformulation strategy. in a relevance feedback cycle, the user is presented with a list of the retrieved documents and, after examining them, marks those which are relevant. In practice, only the top 10(or 20) ranked documents need to be examined. The main idea consists of selecting important terms, or expressions, attached to the documents that have been identified as relevant by the user, and of enhancing the importance of these terms in a new query formulation. The expected effect is that the new query will be moved towards the relevant documents and away from the non-relevant ones.

Local analysis techniques are interesting because they take advantage of the local context provided with the query. In this regard, they seem more appropriate than global analysis techniques. In a local strategy, the documents retrieved for a given query q are examined at query time to determine terms for query expansion. This is similar to a relevance feedback cycle but might be done without assistance from the user.

Key words : Relevance feedback, local context analysis, similar, query expansion, p-norm

I. Introduction

Boolean retrieval is simple and elegant. However, since there is no provision for term weighting, no ranking of the answer set is generated. As a result, the size of the output might be too large or too small[1].

The vector model recognizes that the use of binary weights is too limiting and proposes a framework in which partial matching is possible. This is accomplished by assigning non-binary weights to index terms in queries and in documents. These term weights are ultimately used to compute the degree of similarity between each document stored in the system and the user query.

By sorting the retrieved documents in decreasing order of this degree of similarity, the vector model takes into consideration documents which match the query terms only partially. The main resultant effect is that the ranked document answer set is a lot more precise than the document answer set retrieved by the Boolean model.

Relevance feedback is the most popular query reformulation strategy. in a relevance feedback cycle, the user is presented with a list of the retrieved documents and, after examining them, marks those which are relevant. In practice, only the top 10(or 20) ranked documents need to be examined. The main idea consists of selecting important terms, or expressions, attached to the documents that have been identified as relevant by the user, and of enhancing the importance of these terms in a new query formulation. The expected effect is that the new query will be moved towards the relevant documents and away from the non-relevant ones[2].

Local analysis techniques are interesting because they take advantage of the local context provided with the query. In this regard, they seem more appropriate than global analysis techniques. In a local strategy, the documents retrieved for a given query q are examined at query time to determine terms for query expansion. This is similar to a relevance feedback cycle but might be done without assistance from the user[3].

The approach is based on the use of noun groups, instead if simple keywords, as document concepts. For query expansion, concepts are selected from the top ranked documents based on their co-occurrence with query terms. However, instead of documents, passages are used for determining co-occurrence[4].

II. Extended Boolean Model

The extended Boolean model, introduced in 1983 by Salton, Fox, and Wu is based on a critique of a basic assumption in Boolean logic as follows. Consider a conjunctive Boolean query given by $q=kx \wedge ky$. According to the Boolean model, a document which contains either the term kx or the term ky is as irrelevant as another document which contains neither of them. However, this binary decision criteria frequently is not in accordance with common sense. An analogous reasoning applies when one considers purely disjunctive queries.

When only two terms are considered, we can plot queries and documents in a two-dimensional map as shown in Figure 2.1. A document dj is positioned in this space through the

Manuscript received January 14, 2002 ;
revised February 20. 2002

adoption of weights wx,j and wy,j associated with the pairs [kx, dj] and [ky, dj], respectively. We assume that these weights are normalized and thus lie between 0 and 1. For instance, these weights can be computed as normalized tf-idf factors as follows.

$$w_{x,j} = f_{x,j} \times \frac{idf_x}{\max_i idf_i} \tag{1}$$

where, as defined by equation 1, fx,j is the normalized frequency of term kx in document dj and idfi is the inverse document frequency for a generic term ki.

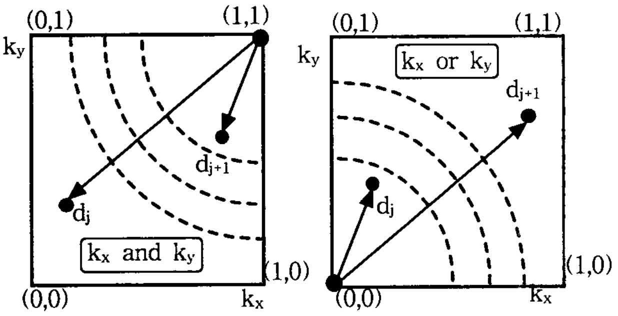

Fig 2.1 Extended Boolean logic considering the space composed of two terms kx and ky only.

For simplicity, in the remainder of this section, we refer to the weight wx,j as x, to the weight wy,j as y, and to the document vector $\vec{d_j} = (w_{x,j}, w_{y,j})$as the point dj=(x,y). Observing Figure 2.1 we notice two particularities. First, for a disjunctive query qor =kx ∨ ky, the point (0,0) is the spot to be avoided. This suggests taking the distance from (0,0) as a measure of similarity with regard to the query qor. Second, for a conjunctive query qand = kx ∧ ky, the point(1,1) is the most desirable spot. This suggests taking the complement of the distance from the point (1,1) as a measure of similarity with regard to the query qand. Furthermore, such distances can be normalized which yields,

$$sim(q_{\text{or}}, d) = \sqrt{\frac{x^2 + y^2}{2}} \tag{2}$$

$$sim(q_{\text{and}}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}} \tag{3}$$

If the weights are all Boolean, a document is always positioned in one of the four corners and the values for sim(qor, d) are restricted to 0, $1/\sqrt{2}$, and 1. Analogously, the values for sim(qand, d) are restricted to 0, 1- $1/\sqrt{2}$, and 1. Given that the number of index terms in a document collection is t, the Boolean model discussed above can be naturally extended to consider Euclidean distances in a t-dimensional space. However, a more comprehensive generalization is to adopt the theory of vector norms as follows.

The p-norm model generalizes the notion of distance to include not only Euclidean distances but also p-distances, where $1 \leq p \leq \infty$ is a newly introduced parameter whose value must be specified at query time. A generalized disjunctive query is now represented by

$$q_{\text{or}} =_{k_1} \vee^p k_2 \vee^p \ldots \vee^p k_m$$

Analogously, a generalized conjunctive query is now represented by

$$q_{\text{and}} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$$

The respective query-document similarities are now given by

$$sim(q_{\text{or}}, d_j) = \left(\frac{x_1^p + x_2^p + \ldots + x_m^p}{m} \right)^{\frac{1}{p}} \tag{4}$$

$$sim(q_{\text{and}}, d_j) = 1 - \left(\frac{(1-x_1)^p + (1-x_2)^p + \ldots + (1-x_m)^p}{m} \right)^{\frac{1}{p}} \tag{5}$$

where each xi stands for the weight wi,d associated to the pair [ki, dj]. The p-norm as defined above enjoys a couple of interesting properties as follows. First, when p=1 it can be verified that

$$sim(q_{\text{or}}, d_j) = sim(q_{\text{and}}, d_j) = \frac{x_1 + \ldots + x_m}{m} \tag{6}$$

Second, when p=∞ it can be verified that

$$sim(q_{\text{or}}, d_j) = \max(x_i)$$
$$sim(q_{\text{and}}, d_j) = \min(x_i) \tag{7}$$

Thus, for p=1, conjunctive and disjunctive queries are evaluated by a sum of term-document weights as done by vector-based similarity formulas. Further, for p = ∞, queries are evaluated according to the formalism of fuzzy logic. By varying the parameter p between 1 and infinity, we can vary the p-norm ranking behavior from that of a vector-like ranking to that of a Boolean-like ranking. This is quite powerful and is a good argument in favor of the extended Boolean model.

The provessing of more general queries is done by grouping the operators in a predefined order. For instance, consider the query q $= (k_1 \wedge^p k_2) \vee^p k_3$ The similarity sim(q, dj) between a document dj and this query is then computed as

$$sim(q, d) = \left(\frac{\left(1 - \left(\frac{(1-x_1)^p + (1-x_2)^p}{2} \right)^{\frac{1}{p}} \right)^p + x_3^p}{2} \right)^{\frac{1}{p}} \tag{8}$$

This procedure can be applied recursively no matter the number of AND/OR operators.

III. Information Retrieval model of RF and LCAF

3.1 Information Retrieval Model of Relevance Feedback

The main idea consists of selecting important terms, or expressions, attached to the documents that have been identified as relevant by the user, and of enhancing the importance of these terms in a new query formulation[4]. The expected effect is that the new query will be moved towards the relevant documents and away from the non-relevant ones[2].

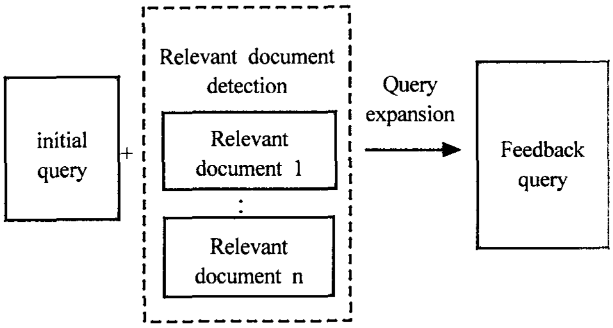

Fig. 3.1 General relevance feedback

The application of relevance feedback to the vector model considers that the term weight vectors of the documents identified as relevant have similarities among themselves. Further, it is assumed that non-relevant documents have term-weight vectors which are dissimilar from the ones for the relevant documents. The basic idea is to reformulate the query such that it gets closer to the term-weight vector space of the relevant documents.

Consider first the unrealistic situation in which the complete set Cr of relevant documents to a given query q is known in advance. In such a situation, it can be demonstrated that the best query vector for distinguishing the relevant documents from the non-relevant documents is given by

$$\vec{q}_{opt} = \frac{1}{|C_r|}\sum_{\forall \vec{d}_j \in C_r} \vec{d}_j - \frac{1}{N-|C_r|}\sum_{\forall \vec{d}_j \notin C_r} \vec{d}_j \quad (9)$$

Dr : set of relevant documents, as identified by the user, among the retrieved documents

Dn : set of non-relevant documents among the retrieved documents

Cr : set of relevant documents among all documents in the collection

|Dr|,|Dn|,|Cr| : number of documents in the sets Dr, Dn, and Cr, respectively.

α, β, γ : tuning constants.

The problem with this formulation is that the relevant documents which compose the set Cr are not known a priori. In fact, we are looking for them. The natural way to avoid this problem is to formulate an initial query and to incrementally change the initial query vector. This incremental change is accomplished by restricting the computation to the documents known to be relevant at that point. There are three classic and similar ways to calculate the modified query $\vec{q}_m$ as follows

Standard_Rocchio :

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|D_r|}\sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|}\sum_{\forall \vec{d}_j \in D_n} \vec{d}_j \quad (10)$$

3.2 Information Retrieval Model of Local Context Analysis Feedback

The local context analysis procedure operates in three steps.

- First, retrieve the top n ranked passages using the original query. This is accomplished by breaking up the documents initially retrieved by the query in fixed length passages and ranking these passages as if they were documents.
- Second, for each concept c in the top ranked passages, the similarity sim(q, c) between the whole query q and the concept c is computed using a variant of tf-idf ranking.
- Third, the top m ranked concepts are added to the original query q. To each added concept is assigned a weight given by 1-0.9 × i/m where i is the position of the concept in the final concept ranking. The terms in the original query q might be stressed by assigning a weight equal to 2 to each of them.

The second on is the most complex and the one which we now discuss.

The similarity sim(q, c) between each related concept c and the original query q is computed as follows.

$$sim(q,\ c) = \prod_{k_i \in q}\left(\delta + \frac{\log(f(c,k_i) \times idf_c)}{\log n}\right)^{idf_i} \quad (11)$$

where n is the number of top ranked passages considered. The function f(c, ki) quantifies the correlation between the concept c and the query term ki and is given by

$$f(c,k_i) = \sum_{j=1}^{n} pf_{i,j} \times pf_{c,j} \quad (12)$$

where $pf_{i,j}$ is the frequency of term ki in the j-th passage and $pf_{c,j}$ is the frequency of the concept c in the j-th passage. Notice that this is the standard correlation measure defined for association clusters but adapted for passages. The inverse document frequency factors are computed as

$$idf_i = \max\left(1, \frac{\log_{10} N/np_i}{5}\right) \quad (13)$$

$$idf_c = \max\left(1, \frac{\log_{10} N/np_c}{5}\right) \quad (14)$$

where N is the number of pasages in the collection, npi is the number of passages containing the term ki, and npc is the number of passages containing the concept c.

The factor δ is a constant parameter which avoids a value equal to zero for sim(q, c). Usually, δ is a small factor with values close to 0.1. Finally, the idfi factor in the exponent is introduced to emphasize infrequent query terms.

Ⅳ. Experimentation and Result

When considering retrieval performance evaluation, we should first consider the retrieval task that is to be evaluated. Consider and example information request I and its set relevant documents Let |R| be the number of documents in this set. Assume that a given retrieval strategy processes the information request I and generates a document answer set A. Let |A| be the number of documents in this set. Further, let |Ra| be the number of documents in the intersection of the sets R and A.

- Recall is the fraction of the relevant documents (the set R) which has been retrieved
- Precision is the fraction of the retrieved documents (the set A) which is relevant

$$\text{Recall} = \frac{|Ra|}{|R|} \qquad (15)$$

$$\text{Precision} = \frac{|Ra|}{|A|} \qquad (16)$$

A single measure which combines recall and precision might be of interest. One such measure is the harmonic mean F of recall and precision which is computed as

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \qquad (17)$$

where r(j) is the recall for the j-th document in the ranking, P(j) is the precision for the j-th document in the ranking, and F(j) is the harmonic mean of r(j) and P(j). The function F assumes values in the interval [0, 1]. It is 0 when no relevant documents have been retrieved and is 1 when all ranked documents are relevant. Further, the harmonic mean F assumes a high value only when both recall and precision are high. Therefore, determination of the maximum value for F can be interpreted as an attempt to find the best possible compromise between recall and precision.

There is experimentation result comparison for P-norm and relevance feedback retrieval in Table 4.1, it is show that relevance feedback retrieval result more improvement 74.59% at recall, 82.5% at precision for initial retrieval result.

Table 4.1 Experimentation result comparison for P-norm and relevance feedback retrieval

division / measure	P-norm	Relevance feedback	Increase rate
Recall	0.362	0.63	+0.27(+74.59)
precision	0.40	0.73	+0.33(+82.50)

Table 4.2 P-norm and relevance feedback recall (document number limit)

division / measure	Recall		
	P-norm	Relevance feedback	Increase rate
document number≤10	0.28	0.48	+0.20(+71.43)
document number≤20	0.47	0.78	+0.31(+65.96)

Table 4.3 P-norm and relevance feedback precision (document number limit)

division / measure	Precision		
	P-norm	Relevance feedback	Increase rate
document number≤10	0.42	0.75	+0.33(+78.57)
document number≤20	0.39	0.71	+0.32(+82.05)

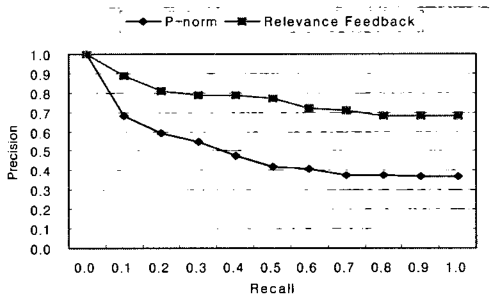

Fig. 4.1 Retrieval experimentation result for P-norm and relevance feedback

A single measure which combines recall and precision might be of interest.

Table 4.4 Single Measure appraisement for P-norm retrieval using harmonic mean

Recall	Precision	Harmonic Mean	Total Harmonic Mean
0.1	0.68	0.170	0.429
0.2	0.59	0.299	
0.3	0.55	0.389	
0.4	0.475	0.434	
0.5	0.42	0.457	
0.6	0.41	0.487	
0.7	0.374	0.487	
0.8	0.374	0.510	
0.9	0.366	0.521	
1.0	0.366	0.536	

Table 4.5 Single Measure appraisement for relevance feedback retrieval using harmonic mean

Recall	Precision	Harmonic Mean	Total Harmonic Mean
0.1	0.89	0.18	0.5758
0.2	0.81	0.327	
0.3	0.79	0.435	
0.4	0.79	0.531	
0.5	0.77	0.606	
0.6	0.72	0.653	
0.7	0.71	0.704	
0.8	0.68	0.735	
0.9	0.68	0.778	
1.0	0.68	0.809	

Fig. 4.1 show that retrieval experimentation result for P-norm and relevance feedback when retrieval document limict 20. This Fig. 4.1 also show that precision wasn't change largely at 0.6 point at recall for all result of P-norm and relevance feedback retrieval

Table 4.6 show that result comparison of relevance feedback(RF) and local context analysis feedback(LCAF) retrieval experimentation, LCAF result is more improve of 3.173% recall, 12.82% precision for relevance feedback retrieval result.

Table 4.6 Result comparison of RF and LCAF retrieval experimentation

division / measure	RF	LCAF	Increase rate
Recall	0.63	0.65	+0.02(+3.174)
Precision	0.73	0.78	0.05(+12.82)

Table 4.7 Recall of RF and LCAF(document number limit)

division / measure	Recall		
	RF	LCAF	Increase rate
document number≤10	0.48	0.52	+0.04(+8.33)
document number≤20	0.78	0.79	+0.01(+1.282)

Table 4.8 Precision of RF and LCAF(document number limit)

division / measure	Precision		
	RF	LCAF	Increase rate
document number≤10	0.75	0.77	+0.02(+2.67)
document number≤20	0.71	0.74	+0.03(+4.23)

Fig. 4.2 show that experimentation of RF and LCAF retrieval when retrieval document limit 20.

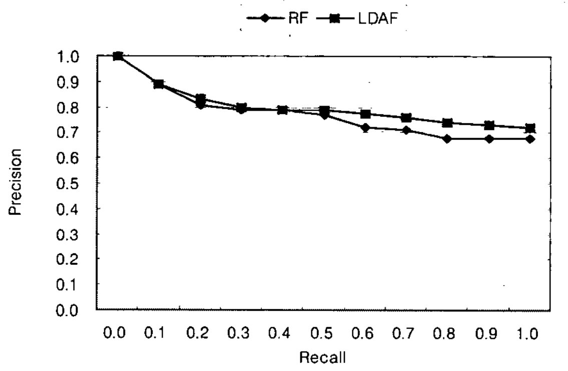

Fig. 4.2 Experimentation result of RF and LCAF retrieval

A single measure which combines recall and precision might be of interest.

Table 4.9 Single measure appraisement of RF using harmonic mean

Recall	Precision	Harmonic Mean	Total Harmonic Mean
0.1	0.89	0.18	0.5758
0.2	0.81	0.327	
0.3	0.79	0.435	
0.4	0.79	0.531	
0.5	0.77	0.606	
0.6	0.72	0.653	
0.7	0.71	0.704	
0.8	0.68	0.735	
0.9	0.68	0.778	
1.0	0.68	0.809	

Table 4.10 Single measure appraisement of LCAF using harmonic mean

Recall	Precision	Harmonic Mean	Total Harmonic Mean
0.1	0.89	0.180	0.5898
0.2	0.831	0.322	
0.3	0.8	0.437	
0.4	0.79	0.531	
0.5	0.79	0.612	
0.6	0.773	0.675	
0.7	0.76	0.728	
0.8	0.74	0.769	
0.9	0.729	0.809	
1.0	0.72	0.837	

Fig. 4.3 show that experimentation of P-norm and RF and LCAF retrieval when retrieval document limit 20.

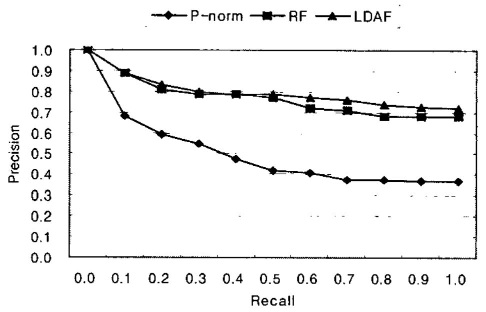

Fig. 4.3 Experimentation result of P-norm, RF, LCAF

V. Conclusion

Local analysis techniques are interesting because they take advantage of the local context provided with the query. In this regard, they seem more appropriate than global analysis techniques. Furthermore, many positive results have been reported in the literature. The application of local analysis techniques to the Web, however, has not been explored and is a promising research direction.

References

[1] A. Bookstein. Implication of Boolean structure for probabilistic retrieval. In Proc. of the 8th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pages 11-17, Montreal, Canada, 1995.

[2] Donna Harman. Relevance feedback revisited. In Proc. of the 5th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1-10, Copenhagen, Denmark, 1992.

[3] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In Proc. ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 4-11, Zurich, Switzerland, 1996.

[4] Baeza-Yates, R. and Ribeiro-Neto, Berthier. Moderm Information Retrieval, addison-wesley Pub. Co(sd), 1992.

Young-Chon Kim

1992 : Dept. of Computer Science, Kwangju University(B.S)

1996 : Dept. of Computer Engineering, Chosun University(M.S)

1998~now : Dept. of Computer Science, Chosun University Doctoral Student

Research Interests : Software engineering(Reuse, metrics), Object-oriented software(metrics), Electronic Commerce, Information Retrieval.

Sung-Joo Lee

1970 : Dept. of Physics Sciences, Hannam University(B.S)

1992 : Dept. of Computer sciences, Kwangwoon University(M.S)

1998 : Dept. of Computer sciences, Catholic University of Daegu(Ph.D)

1988~1990 : Chief, Computer Center, Chosun University

1995~1997 : President, Information Science, Chosun University

1981~now : Professor in the Dept. of Computer Engineering, Chosun University

Research Interests : Software engineering, Programming Language, Object-oriented software, Rough set.