Effect of Filter Size on Fusion Function in Information Retrieval

Nagammapudhur Gopalan [1] and Krishnan Batri [2]
[1]Department of Computer Applications, National Institute of Technology, India
[2]Department of Computer Science and Engineering, National Institute of Technology, India

Abstract: *An attempt to detach the worst performing schemes for improving the effectiveness of the information retrieval system, which combines the multiple sources of evidence, has been presented in this paper. The disturbances caused by the ill performing retrieval schemes are studied by using the concept of filter. The size of the filter is altered to examine the efficiency of the retrieval system. The experiments are conducted over the three-benchmark test collections viz., ADI, CISI and MED. The results indicate that, the efficiency of the retrieval system varies with the filter size and the maximum improvement in performance is achieved only at particular filter size, which is different for various fusion functions.*

Keywords: *Information retrieval, fusion, filter, overlap, and precision*.

Received December 6, 2006; accepted February 21, 2007

1. Introduction

Information Retrieval (IR) task involves the selection of pertinent documents among a collection of document (corpus) like web, digital library etc., [14, 19, 10]. According to the relevance score, the articles are arranged in descending order by the retrieval scheme, which is employed in the IR system. The relevance score is calculated based on the degree of match or map between the users' specified keywords (query) and the stored index terms of the corpus. Performance of the retrieval systems may be evaluated using the precision and recall measures [19].

Retrieval schemes, the loci-classicus of IR, have varying performance and warranted a thorough look at a new concept called fusion associated with them. Fusion is the methodology of combining multiple data sources. In IR, it combines the various retrieval schemes and utilizes the following: (1) skimming effect, (2) chorus effect and (3) dark horse effect. The schemes used to collect relevant documents arrange them in various order of their importance. When retrieval is made using a combination of schemes, if the top ranking documents under each of the list are selected, then the phenomenon is termed as *skimming effect*. The *chorus effect* assigns a high degree of relevance to the documents found in a majority of lists of the relevant documents returned by the retrieval schemes. Subsequently, these highly relevant documents are deemed to be the final relevant list corresponding to the fusion of the retrieval schemes. The *Dark horse effect* is one in which a retrieval approach may produce some of the items with unusually accurate (or inaccurate) estimates of relevance score.

This paper focuses on the study of the effect of filter size on fusion functions. Variation in performance of the fusion functions is recorded at different filter size. In future, it is planned to develop an algorithm, which effectively nullify the disturbances caused by the ill performing schemes, based on the results obtained in this paper.

2. Data Fusion

The fusion function, which combines the multiple data sources, is based on either the basic set theoretic operations like union, intersection or normal arithmetic operations. Data fusion finds an extended application to a wide variety of scientific and engineering areas like remote sensing, robotics, surveillance etc. It has been observed that the fusion function based on the Chorus effect yields better performance in comparison with others [11, 12].

2.1. Chorus Effect

'Two heads are better than one' is the basic notion upon which the concept of fusion is constructed. This is synonymous to the voting scheme, where a winning candidate acquires the majority of polled votes. In IR, a document is more likely to be relevant, when more than one retrieval schemes suggest, that it is relevant. Under a fusion function, the chorus effect got amplified, when few of the retrieval schemes return very low relevance score and the rest return high scores. These low score-returning schemes act as noise; disturbing the information-bearing signal. Hence, by effectively filter out the noises, reduction in

the amplification of the chorus effect is achieved and it may lead to improvement in performance. In this case, the design of the filter for the filtering process becomes a critical task.

3. Fusion Techniques

Fisher [4] made an early attempt to improve the effectiveness of the IR system by merging two boolean searches together. One of the searches operated on the title word while the other is used to explore the manually generated index terms. This method is confined only to two sources. A linear combination method for combining multiple sources by assigning weights to the individual schemes was studied by Belkin and Croft [3, 1]. The final relevance score of a document assigned by the weighted linear combination method is given by

$$R(q,d) = \sum_{i=1}^{k} \theta_i . E_i(q,d) \quad (1)$$

where θ_i = weight of the i^{th} retrieval scheme. $E_i(q, d)$ =Relevance score returned by the i^{th} retrieval scheme and k = Number of retrieval schemes to be fused.

The weighted linear combination method has the limitation of requiring prior knowledge about the retrieval systems to assign the weights [16]. The Comb-functions for combining scores have been proposed by Fox and Shaw. The various comb-functions used for combining scores are shown in Table 1 [5] and [6]. Lee conducted extensive work on Comb-functions and proposed new rationales, indicators for data fusion [11, 12, 13].

The training data for the fusion operation are used to select the best functioning scheme with appropriate weight. Probabilistic approach is used for this purpose. The best performing scheme is selected automatically from the pool of schemes in spite of the appreciable performance of the others. In order to overcome this drawback, Bilhart proposed a heuristic based data fusion algorithm, which uses Genetic Algorithm (GA) [2]. His algorithm assigns weights to independent retrieval schemes and selects the significant ones for fusion.

Table 1. Comb-functions for combining scores.

Comb-functions	Explanation
Comb-MIN	Minimum of all relevance scores
Comb-MAX	Maximum of all relevance scores
Comb-SUM	Summation of all relevance scores
Comb-ANZ	Comb-SUM ÷ Number of non zero relevance scores
Comb-MNZ	Comb-SUM × Number of non zero relevance scores

Since the fusion process combines the results from multiple sources; the number of sources participating in the merging operation determines the performance of the combination function. Vogt varies the number of sources participating in the fusion and conclude with some new remarks [17].

The proposed approach is entirely different from Vogt. In our approach, the retrieval schemes, which are returning low relevance scores to a document, are not considered for the fusion and the scores returned by them are completely filtered out. This may vary from one document to another. Hence, the fusion function considers the contribution of a scheme for a particular document alone and it may not be the same for some other documents. Where in vogt approach, the contribution of the retrieval scheme may or may not considered for entire documents in the corpus.

4. Information Content of Retrieval Schemes

When all retrieval schemes return high or equal relevance scores the merging process becomes less involved [18] and hence it is desirable. The certainty about the relevance of the documents as indicted by the score may be analysed by using the statistical information theory [15].

The retrieval schemes assign score to documents in order to provide information about their relevance as in the case of symbols in the statistical information theory. Hence, the participating scheme may be considered as message symbols for further analysis. As the combination function operates on all retrieval schemes, it may be considered as a message source. Let 's' be the message symbol and the probability of their occurrence be 'p'. The occurrences of the symbols are treated as independent events. If there are n IR schemes, s and p become

$$s = \{ s_1, s_2, s_3, ..., s_n \} \quad (2)$$

$$p = \{ p_1, p_2, p_3, ..., p_n \}, \quad with \quad \sum_{i=1}^{n} p_i = 1 \quad (3)$$

The probability of occurrences of the message symbols may be calculated from the relevance scores R_j $(1 \le j \le n)$ as

$$p_j = \frac{R_j}{\sum_{i=1}^{n} R_i} \quad (4)$$

Hence, it may be inferred that

$$p_i \propto R_i$$

$$p_i \propto \frac{1}{\sum_{i=1}^{n} R_i} \quad (5)$$

Let the j^{th} retrieval scheme assign a maximal score to a particular document. Equivalently, the message symbol j has a high probability of occurrence. Further,

$$I(j) = -\log(p_j) \quad (6)$$

where $I(j) \to 0$, $P(j) \to 1$ and $I(j) \to 1$, $P(j) \to 0$.

The desired condition for a high probability to a symbol leads to a very low information content (since the information content of the highly probable event is nil). Also when the scores are equal

$$\forall i, \sum_{i=1}^{n} R_i = n.R_i \quad (7)$$

$$p_i = \frac{R_i}{\sum_{i=1}^{n} R_i} = \frac{R_i}{n.R_i} = \frac{1}{n} \quad (8)$$

The entropy is used as the performance indicator for analysing the characteristics of the message source and it is given by

$$H(Y) = -\sum_{i=1}^{n} p_i . \log p_i \quad (9)$$

When the probability of occurrence of all message symbols is equal, the entropy of the source becomes

$$H(Y) = \log(n) \quad (10)$$

In view of the statistical communication theory, the desired criteria for the fusion may be restated as

- The information content of the message symbol should be minimum and
- The message source ought to have the maximum entropy.

When probabilities of the symbols are unequal and the symbol *i* has the maximum probability in comparison with the others.

$$p_i \neq p_j \neq ..... \neq p_n \ \ and \ \ p_i > p_j, \ i \neq j \ \forall j. \quad (11)$$

consequently,

$$I(i) \neq I(j), \forall j, i \neq j \quad (12)$$

$$H(Y) \neq \log(n) \quad (13)$$

The desired condition may be achieved by increasing the probabilities of message symbols by deleting the low relevance scores in the denominator of equation 4. If $\sum_{k=1}^{m} R_k$ is the sum of '*m*' low relevance scores to be deleted, then the probability of the message symbol '*I*' becomes

$$p_i = \frac{R_i}{\sum_{j=1}^{n} R_j - \sum_{k=1}^{m} R_k} \quad (14)$$

When the low relevance scores are deleted one by one, $P_i \rightarrow 1$, $Ii \rightarrow 0$. This is the unwanted side effect. The number of low relevance scores '*m*' to be deleted may play a vital role in meeting the desired conditions and the concept of filter is used to determine them.

5. Effect of Filter Size on Fusion Functions

Filter is a device or mechanism used to filter out the noise from the desired ones. Filter allows the signal which lie above the cutoff frequency. The signal is usually expressed in decibels and for a signal with frequency a_1, an equivalent may be calculated as

$$20*log_{10}a_1 \quad (15)$$

The magnitude of the relevance score of a document determines it's significance. In the proposed study, the filter size is varied to fix the range of relevance score for the participating retrieval schemes. The change in performance of the fusion function is recorded at different filter size. There are two possible ways to test the performance variation:

- Upper edge of the filter is fixed at the universal maximum (which is 1 in our experiment as the relevance score of the documents are in the range of (0-1)) and the filter size is continuously varied.
- The upper edge as well as the filter size is varied. The upper edge of the filter is fixed at the maximum relevance score of a particular document, which is different for different queries and documents.

In the experiment, the second one was chosen. The number of relevant scores present inside the filter is treated as the *overlap* value and the scores that lie outside the filter are deleted.

5.1. Fusion Functions

The *Comb-MAX*, *Comb-MNZ* and *Comb-SUM* functions are selected for the study. In *Comb-MAX*, the maximum value remains the same as the filter always retains it. Hence, in order to carryout the study, the maximum relevance score is multiplied with newly calculated overlap value. In *Comb-MNZ*, the existing overlap value is being replaced with the newly calculated one. In *Comb-SUM*, only the values present inside the filter are considered. Figure 1 shows the filter based *Comb-functions* (i.e., *F-Combfunctions*) used in the study.

F-CombMAX	Maximum of all relevance score × γ
F-CombSUM	Sum of all relevance Scores lie inside the filter
F-CombMNZ	Sum of all relevance Scores lie inside the filter × γ

Where,

γ - The number of relevance scores present inside the filter

Figure 1. F-Combfunctions.

F-CombSUM and *F-CombMNZ* functions linearly combine the relevance scores and by doing so these two functions extract the advantages of the chorus effect. The *F-CombMAX* utilizes the advantages of both skimming and chorus effect.

5.2. Data Collections and Retrieval Schemes

The experiment is conducted over the three benchmark test document collections viz., MED [9], CISI [8], and ADI [7].

Table 2. Characteristics of datasets.

	ADI	CISI	MED
Number of documents	82	1460	1033
Number of terms	374	5743	5831
Number of queries	35	35	30
Average number of document relevant to a query	5	8	23
Average number of terms per document	45	56	50
Average number of terms per query	5	8	10

Table 2 shows the characteristics of the data sets. The uniform environment is being maintained by using the same stop word list (smart stop word list), stemmer algorithm (porter stemmer) and same weight assignment mechanism. The Term Frequency and Inverse Document Frequency (TF IDF) weight assignment method is used and are given in equations 15 and 16.

$$w_t = log_{10}\left(1 + \frac{N}{f_t}\right) \quad (15)$$

$$w_{d,t} = f_{d,t} \cdot w_t \quad (16)$$

where N = total number of document in the corpus, f_t= number of documents containing the term t, w_t= term weight, $w_{d,t}$= document term weight, $f_{d,t}$ = frequency of the term t in document d.

The similarity measures of Vector Space Model (VSM) and *P-norm* model with P values 1.5, 2.5 and 3.5 are chosen as retrieval schemes. The similarity measures of VSM are given by

cosine similarity measure

$$S(q,d) = \frac{\sum_{t \in q \cap d} (w_{q,t} * w_{d,t})}{W_q * W_d} \quad (17)$$

inner product

$$S(q,d) = \sum_{t \in q \cap d} w_{q,t} * w_{d,t} \quad (18)$$

dice coefficient

$$S(q,d) = \frac{2 \sum_{t \in q \cap d} w_{q,t} * w_{d,t}}{W_q^2 + W_d^2} \quad (19)$$

jaccard coefficient

$$S(q,d) = \frac{\sum_{t \in q \cap d} (w_{q,t} * w_{d,t})}{W_q^2 + W_d^2 - \sum_{t \in q \cap t} (w_{q,t} * w_{d,t})} \quad (20)$$

where $S(q,d)$ = similarity score of a document d with respect to query q, $w(q,t)$ = weight of the term t in the query q, $w(d,t)$ = weight of the term t in the document d, W_q = weight of the query q and W_d = weight of the document d.

The conjunctive query form of *P-norm* model is also used as a retrieval scheme in the experiment and it is shown in equation 21.

$$Sim(q_{and}, d_j) = 1 - \left(\frac{(1-w_1)^p + (1-w_2)^p + \ldots\ldots + (1-w_m)^p}{m} \right)^{1/p} \quad (21)$$

where w_m is the weight of the m^{th} index term and $1 \le p \le \infty$.

The scores returned by the various retrieval schemes based on the weight of the index terms are of various ranges. The fusion process may be dominated by the scheme, which returns score of higher range. In order to maintain a uniform environment, normalization is used. In the experiment `*Max normalization*' is selected for this purpose and it is given below.

$$R_{normalized} = \frac{R_{unnormalized}}{R_{max}} \quad (22)$$

where $R_{unnormalized}$ is the relevance score returned by a retrieval scheme and R_{max} is the maximum relevance score returned by a retrieval scheme.

5.3. Variation in Filter Size: An Analysis

The effect of varying the filter size on fusion function is analysed using the 11-point interpolated precision [19]. The average value of the 11-point interpolated value for the *CombMNZ* over the three-document collection at different filter size is shown in Figure 2.

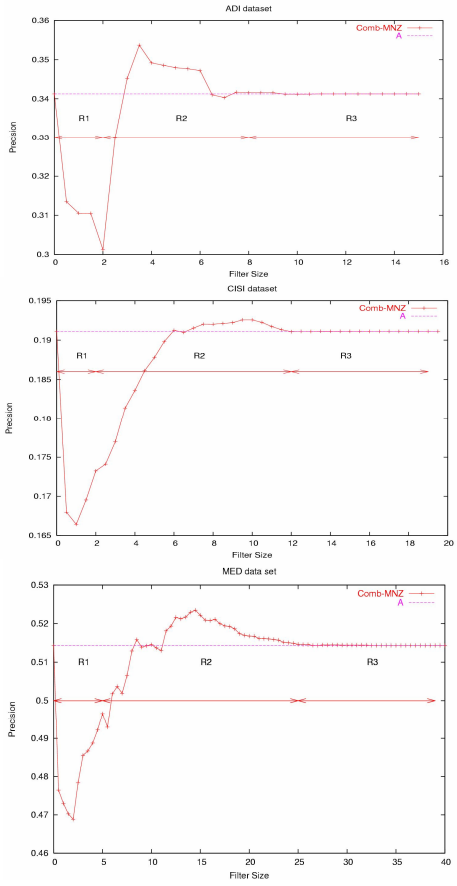

Figure 2. Performance of *CombMNZ* functions at various filter size.

The line marked as 'A' in the graph is used as the reference line (the relevance score at 0 dB; 100%) for comparison purpose. In the graph, at 0 dB the performance of the functions is recorded as such without imposing the filter. The size of the filter is varied in steps of 0.5 dB. The upper edge of the filter for a document is fixed at the maximum of all relevance scores (maximum of all relevance scores returned by the seven retrieval schemes discussed in previous subsection and it is converted in to decibels). The lower edge is varied.

$$\textit{Lower edge} = \textit{upper edge} - \textit{chosen step size of the filter} \quad (23)$$

for example if the upper edge is 1 dB and the chosen step size is 0.5 dB then filter size is 0.5 dB. The scores, which lie inside the filter, are considered for fusion. The performance of the other *Comb-functions* (namely *Comb-MAX* and *Comb-SUM*) are qualitatively same, hence they are not shown separately. The performance of the fusion functions is analysed by dividing the graphs in to three regions R_1, R_2 and R_3.

- *Region* R_1: In this region, size of the filter is very small and the fusion functions concentrate only on the retrieval schemes that return higher relevance score. Hence, only very few retrieval schemes are considered for fusion and the remaining become unused. This leads to degradation in performance.
- *Region R2*: Filter size for this region is moderate. The number of schemes to be considered for fusion are neither too small nor too many. This leads to the improvement in performance.
- *Region R3*: In this region, almost all retrieval schemes participating in the fusion are considered and this ends up with the amplification of chorus effect. As a result, the performance of the fusion functions starts to degrade as we gradually move toward right in the graph. The point at which the region R_2 and R_3 meet is termed as *flattening point*.

The precision value at the 0 dB and at the flattening point is quantitatively same. This is due to the fact that at 0 dB there is no filter and as the filter size is increased gradually, at the flattening point all retrieval schemes are included (equivalently no filter is imposed).

5.4. Performance Comparison

The *F-Combfunctions* of the proposed study are compared with the *comb-functions* and the precision values are given in the table 3. The performance has been observed to improve up to a maximum of 13.2% and an average of 3.69%. The filter size, which is responsible for the performance improvement varies from function to function and corpus to corpus. So, an optimal filter size is to be determined for enhancing performance.

Table 3. Comparison of F-Combfunctions and Comb-functions.

CombMAX			
Collection	F-Comb	Comb	% of improv
MED	0.5167	0.454	13.206
ADI	0.3622	0.3475	4.2426
CISI	0.1937	0.1901	2.8270
CombMNZ			
Collection	F-Comb	Comb	% of improv
MED	0.5234	0.5143	1.7159
ADI	0.3537	0.3412	3.6484
CISI	0.1925	0.1911	2.5351
CombSUM			
Collection	F-Comb	Comb	% of improv
MED	0.5228	0.5143	1.6539
ADI	0.3518	0.3411	0.31141
CISI	0.1917	0.1911	0.3287

5.5. Generalized Characteristic Curve

A generalized curve enveloping the effects of filter size is shown in Figure 3.

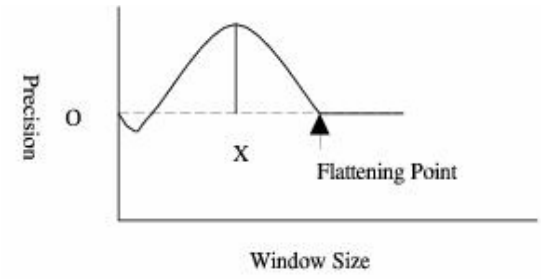

Figure 3. Generalized characteristic curve of effect filter size on fusion functions.

6. Conclusion

The statistical communication theory indicates that the deletion of low relevance scores improve the performance of the fusion functions. Effect of filter size on fusion function is analysed. The *F-CombMAX* achieves significant improvement over the others and hence it may be advantageously used for IR. As *F-CombMAX* utilizes the advantages of both skimming and chorus effect, where as, the *F-CombSUM* and *F-CombMNZ* confined to chorus effect alone. The performance improvement is achieved at a particular filter size, which is different for different document. Hence, it is necessary to find a universal filter size that suits all types of documents. In future, it is planned to develop a new algorithm for the effective filtering based on these results.

References

[1] Belkin N., Cool C., Croft W., and Callan J., "The Effect of Multiple Query Representations on Information Retrieval System Performance," *in Proceedings of the 16th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 339-346, 1993.

[2] Billhart H., "Learning Retrieval Expert Combinations with Genetic Algorithms," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 11, no. 3, pp. 87-113, 2003.

[3] Croft B., "Combining Approaches to Information Retrieval," *in Croft W. (Editor), Advances in Information Retrieval, Chapter 1*, pp. 1-36, Kluwer Academic Publishers, 2000.

[4] Fisher H. and Elchesen D., "Effectiveness of Combining Title Words and Index Terms in Machine Retrieval Searches," *Nature*, pp.109-110, July 1972.

[5] Fox E. and Shaw J., "Combination of Multiple Searches," *in Proceedings of the Second Text Retrieval Conference (TREC-2)*, pp. 243-252, 1994.

[6] Fox E. and Shaw J., "Combination of Multiple Searches," *in Proceedings of the Third Text Retrieval Conference (TREC-3)*, pp. 105-108, 1995.

[7] Korfhage R., *Information Storage and Retrieval*, Willey Computer Publishing, 1997.

[8] Lee J., "Analyses of Multiple Evidence Combination," *in Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267-276, 1997.

[9] Lee J., "Combining Multiple Evidence from Different Properties of Weighting Schemes," *in Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 180-188, 1995.

[10] Lee J., "Combining Multiple Evidence from Different Relevance Feedback Network," *in Proceedings of the 5th International Conference on Advanced Data Base Application*, Melbourne, Australia, pp. 421-430, April 1997.

[11] Salton G. and McGill M., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

[12] Shanon C., "Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.

[13] Vogt C., "Adaptive Combination of Evidence for Information Retrieval," *PhD Thesis*, University of California, San Diego, 1999.

[14] Vogt C., "How Much More is Better? Characterizing the Effects of Adding More IR Systems to a Combination," *in Proceedings of the Conference on Content-Based Multimedia Information Access (RIAO)*, pp. 457-475, 2000.

[15] Yager R., "Fusion of Ordinal Information Using Weighted Median Aggression," *International Journal of Approximate Reasoning*, vol. 18, no. 1, pp. 35-52, 1998.

[16] Yates R. and Neto B., *Modern Information Retrieval*, Pearson Education, 1999.

Nagammapudhur Gopalan received the MSc from Madras University in 1978 and the PhD in applied mathematics from Indian Institute of Science, Bangalore, in 1983. Currently, he is a professor in the Department of Computer Applications in the National Institute of Technology Tiruchirapalli, Tamil Nadu, India. His research interests include algorithms, combinatorics, data mining, and distributed parallel and grid computing.

Krishnan Batri received the ME from Madurai Kamarj University in 2003. Currently, he is a research scholar with the Department of Computer Science and Engineering in the National Institute of Technology Tiruchirapalli, Tamil Nadu, India. His research interests include information retrieval, data fusion, and genetic algorithms.