Kuwait J. Sci. **44 (2)** pp. 78-86, 2017

An effective selection of retrieval schemes for data fusion

Krishnan Batri[1*]

Dept. of Electronics and Communication Engineering
PSNA College of Engineering and Technology, Dindigul
krishnan.batri@gmail.com

Abstract

Merging the results from more retrieval systems/schemes may enhance the performance of the Information Retrieval system. The success of the fusion lies in the selection of the member schemes. This paper explores an effective selection algorithm, which is derived from the filter concept, by treating low-score returning schemes as noises. The proposed algorithm is tested over the three benchmark test collections namely, American Documentation Institute (ADI), Centre for Inventions and Scientific Information (CISI), and Medlars (MED). The consistency of the computed result is tested by paired student-t test. It is observed that the presented algorithm results in significant improvement over the existing combination functions. The improvement in performance of the projected method is due to the reduction in amplification chorus effect caused by the low score returning schemes.

Keywords: Data fusion; filter; information retrieval; overlap; precision; student-t test.

1. Introduction

Information retrieval (IR) is the process of selecting relevant documents from a collection like web, digital library (Salton & McGill, 1986; Yates & Neto, 1999; Korfhage, 1997) etc., known as corpus, using certain IR strategies. Based on the match between the users specified key words or queries and the index terms in corpus, relevant documents are arranged in the descending order of their relevance for retrieval. 'Precision' and 'Recall' are the two measures used to infer the significance of an IR scheme under these circumstances (Yates & Neto, 1999).

Performance of IR schemes varies over different corpus (Wu, 2012). To enhance it, IR schemes are usually combined or fused in a judicial way (Sanke, 2015). Selected documents have been observed to reveal i) skimming effect, ii) chorus effect and iii) dark horse effect in their selection (Croft, 2002). Skimming effect is the selection of top ranking documents under each of the individual IR schemes participating in fusion, while retrieval of documents due to an unexpected key-word match resulting in unusually accurate relevance score estimation is called the dark horse effect. The chorus effect assigns a high degree of relevance to the documents found in a majority of lists of relevant ones returned by the schemes. Consequently, these are deemed to be the final relevant list retrieved by the fusion of IR strategies. The extent of chorus effect amplification depends on the number of low-score returning IR strategies for a selected document and hence need to be filtered out, treating them as noises.

The present paper examines the chorus effect amplification and efficiency of retrieval schemes using filters of various sizes. The low score returning strategies are treated as noises. As they create an illusion about the relevance of the documents and become the cause of degradation in performance. The IR schemes themselves may be considered to be symbols which may be found in a 'message' provided by the fusion function. These assumptions allow the study with in the perspective of information theory (Shannon, 1948) suggesting the use of information content and entropy as performance indicators for the IR schemes and the fusion function used. Both chorus and skimming effects are effectively tapped by a newly proposed fusion function (F-CombMax) resulting in a phenomenal performance improvement vis-a-vis the existing fusion functions found in the literature, which had always shown one of the aforesaid effects percolated in retrieved pages. A paired student-t test applied to the relevant populations of retrieved documents obtained with and without filters has shown that the proposed method is an effective one.

The variation of performance with the existing schemes is analyzed using paired student-t test over different document populations under each of the fusion functions in the proposed set.

2. Prior work in data fusion

The information retrieval process adapts the statistical concepts for searching and retrieving relevant information (Salton & McGill, 1986). As it uses the statistical concepts, the IR is used in machine learning also (Ullah *et al.,* 2016). As various concepts are available, each one of them suffers its own drawback (Ponte & Croft, 1998). Hence extracting the best performance from the existing IR methods becomes a tedious task.

Data fusion for information retrieval was employed by Fisher & Elchesen (1972) by combining two Boolean searches together, one on the title words and the other on the manually generated index terms. A linear combination method for fusing multiple sources by assigning weights to the individual schemes was studied by Croft (2002) and Belkin *et al.* (1994) with the limitation of requiring prior knowledge of the retrieval systems for assigning the weights (Vogt, 1999). The 'CombFunctions' for combining scores that treat all schemes equally have been proposed by Fox & Shaw (1994, 1995). Extensive work on CombFunctions has been carried out by Lee (1995, 1997a,b) proposing new rationales and indicators for data fusion. Using a probabilistic approach, the training data for the fusion operation are used to select the best functioning scheme with appropriate weights (Wu *et al.,* 2014). The scheme with best performance is selected automatically from the pool of schemes in spite of the appreciable performance of the remaining ones. This was overcome by Bilhart (2003), who proposed a heuristic data fusion algorithm that uses Genetic Algorithm (GA) for combining the retrieval scores. The heuristic based method needs some training. The successes of the heuristic based data fusion methods are entirely dependent on the history and training (Ghosh *et al.,* 2015). Some of the CombFunctions, which are used in the present study are shown in Table 1.

Table 1. CombFunctions for combining scores

CombFunctions	Document selection Criterion
CombMAX	Maximum of all relevance scores
CombSUM	Summation of all relevance scores
CombANZ	CombSUM ÷ Number of non zero relevance scores

3. Contribution of the retrieval schemes

The certainty about the relevance of the documents as indicted by the score may be analyzed using the statistical information theory (Shannon, 1948), as it is possible to establish an abstract correspondence between it and the CombFunctions by considering the individual IR strategies participating in a selection to be symbols constituting a message whose source being the fusion function. Let 's' and 'p' be the sets of message symbols and if there are 'n' IR schemes these become

$$s = \{s_1, s_2, \ldots, s_n\} \tag{1}$$

$$p = \{p_1, p_2, \ldots, p_n\} \; with \; \sum_{i=1}^{n} p_i = 1,$$

$$and \; p_j = \frac{R_j}{\sum_{l=1}^{n} R_l} \; (1 \le j \le n) \tag{2}$$

Let the jth retrieval scheme assign a maximal score to a particular document which means that the message symbol j has a high probability of occurrence. Further,

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

Where I (j) → 0 as P (j) → 1 and I (j) → ∞ when P (j) → 0.

The desired condition for a high probability to a symbol leads to a very low information content. The entropy may be used as the performance indicator for analyzing the characteristics of the message source and is given by

$$H(j) = \sum_{i=1}^{n} p_i \,.\log(p_i) \tag{3}$$

When the occurrence of all message symbols is equally likely, the entropy can be written as H (j) = log (n).

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

1. *The information content of the message symbol should be minimum and 2. The entropy of the message source should be maximum.*

Consider a situation where the probabilities of symbols are unequal and the probability of one of them, say p_i is maximum

$$p_1 \neq p_2 \neq ..... \neq p_i \neq p_j \neq p_n \text{ and } p_i > p_j \ \forall j$$

Consequently, H (j) $\neq$ log(n).

The desired condition may be achieved by increasing the probabilities of message symbols by deleting the low relevance scores in the denominator of (2). If $\sum_{k=1}^{n} R_k$ is the sum of 'm' low relevance scores to be deleted, then the probability of the message symbol 'i' becomes

$$p_i = \frac{R_i}{\sum_{j=1}^{n} R_j - \sum_{k=1}^{m} R_k} \quad (4)$$

When the low relevance scores are discarded one by one, $p_i \rightarrow 1$, I (j_i) and H (j) $\rightarrow$ 0, which is the unwanted side effect. The number of low relevance scores 'm' deleted along with their corresponding IR schemes may play a vital role in meeting the desired conditions and the concept of filter is used to determine them.

4. Selection of retrieval schemes

This paper focuses on the concept of filter for selecting the best retrieval schemes. Filters allow the signals above a fixed (range of) cut off frequency. The signal is usually expressed in decibels and for a given signal with frequency (score returned by an IR strategy) λ, its decibel equivalent is given by 20 x $\log_{10}$ λ. Its size can be varied by fixing one of its ends at the maximal score of a document and varying the other end to any specified level. The number of relevant scores present inside the filter is treated as the overlap value (γ) and the scores that lie outside are deleted.

A set of modified fusion function that works with in the filter and the criteria used for selection of documents is defined as follows:

I. F-CombMAX : Maximum relevance score ×γ

II. F-CombSUM : Sum of all relevance Scores lay inside the filter

III. F-CombMNZ : F-CombSUM ×γ

F-CombSUM and F-CombMNZ functions linearly combine the relevance scores and get influenced by the chorus effect whereas the F-CombMAX considers all schemes equally; manifesting the skimming effect.

4.1 Data Collection and retrieval schemes

The experiment is conducted over the three-benchmark test document collections namely: (i) MED (ii) CISI and (iii) ADI under a uniform environment consisting of the same Smart stop word list; Porter's -Stemmer algorithm; and weight assignment. The Table 2 shows the characteristics of these three data sets.

Table 2. Characteristics of the data sets

	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

The Term-Frequency and Inverse-Document Frequency (TF-IDF) weight assignment method is used and the corresponding term- weight (w_t), and document-term weight ($w_{d,t}$) are given by

$$w_t = log_{10}(1 + \frac{N}{f_t}) \quad (5)$$

$$w_{d,t} = f_{d,t} \cdot w_t \quad (6)$$

N = total number of document in the corpus,

f_t = number of documents containing the term t and

$f_{d,t}$ = frequency of the term t in document d.

The similarity measures of Vector Space Model

(VSM) and P-Norm model with P value 1.5, 2.5 and 3.5 are chosen as retrieval schemes (Yates &.Neto,1999). The similarity measures of VSM are given by

$$\frac{\sum_{i\in q\cap d} w_{q,t} * w_{d,t}}{W_q * W_d} \quad (7)$$

$$Inner\ product\ S(q,d) =$$

$$\sum_{i\in q\cap d} w_{q,t} * w_{d,t} \quad (8)$$

$$Dice\ Coefficient\ S(q,d) =$$

$$\frac{2.\sum_{i\in q\cap d} w_{q,t} * w_{d,t}}{W_q^2 + W_d^2} \quad (9)$$

$$Jaccard\ Coefficient\ S(q,d) =$$

$$\frac{\sum_{i\in q\cap d} w_{q,t} * w_{d,t}}{W_q^2 + W_d^2 - \sum_{i\in q\cap d} w_{q,t} * w_{d,t}} \quad (10)$$

Where,

S(q,d) = similarity score of document d with respect to query q,

w(q,d) = 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 and

W_d = weight of the document d.

The conjunctive query form of P-norm model given by

$$Similarity\ S(q_{and}, d_j) = 1 -$$

$$\left(\frac{(1-w_1)^p + (1-w_2)^p + \cdots \ldots + (1-w_m)^p}{m}\right)^{\frac{1}{p}} \quad (11)$$

Where,

w_m = weight of the m^{th} index term and $1 \leq p \leq \infty$.

It is to be noted that w_m = weight of the m^{th} index term and $1 \leq p \leq \infty$.

4.2 Effect of filter size

The effect of varying the filter size on fusion functions in steps of 0.5 dB is analyzed using the 11-point interpolated precision (Korfhage, 1997). The average value of the 11-point interpolated value for the CombMNZ over the three test document collections is shown in the Figure 1.

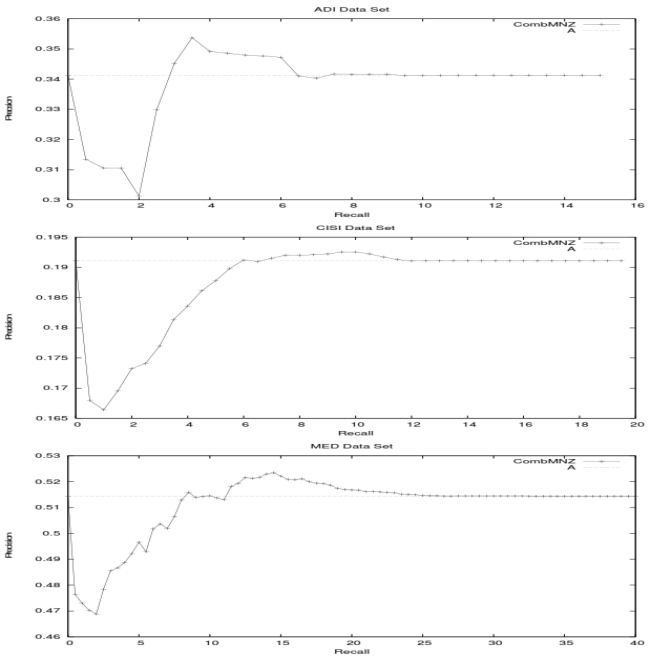

Fig. 1. Performance of the fusion functions at various filter size

The line marked as A in the graph is the reference line (the relevance score at 0 dB; 100%) used for comparison. In the graph at 0 dB, the performances of the functions are recorded as such without imposing the filter. *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 no filter is applied and as the filter size is increased gradually, at the flattening point all retrieval schemes are included (equivalently no filter is imposed).* The performance of the CombMAX and the CombSUM functions are qualitatively same and hence not shown separately.

4.3 Performance comparison

The F-CombFunctions of the proposed study is compared with the CombFunctions and the overall average precision values are given in the Table 3.

Table 3. Comparison of F-CombFunctions and CombFunctions

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

The performance has been found to improve to a maximum of 13.2% and average of 3.69%. The filter size 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.

5. Optimal filter size

The results of the previous section are used to develop an algorithm, which filters out the worst performing schemes. A generalized curve enveloping the effects of filter size is shown in Figure 2.

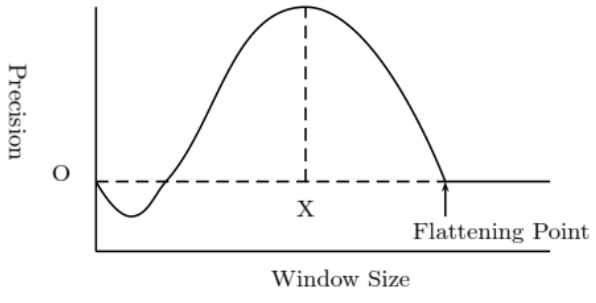

Fig. 2. Generalized characteristic curve of filter effect

The maximum difference among all relevance scores for any generic document at 0 dB gives the filter size at the flattening point, since the precision value at 0 dB and flattening point are same. This significant conclusion is used for computing the size of the filter at the flattening point. As the peak precision value occurs at a point X, which lies below the calculated filter size (below the flattening point), it is necessary to compute the value of X.

5.1 Computing the value of OX

The computed filter size reduced in steps of 0.1 (0.9 times the filter size 0.8 times filter size and so on) and the performance at each filter size is recorded. The experiments are conducted by varying the number of retrieval schemes starting form 2 and ends with 7. The average among all combination is considered for examination purpose. After calculating the overall results, it is planned to test whether the filter size has some significant impact over the performance. ANOVA table is used for this purpose. The hypotheses used in the ANNOVA table are given below.

H0: There is no significant difference among precision value at various filter size.

H1: There is significant difference among precision value at various filter size.

The computed F value is shown in the Table 4.

Table 4. F - Value

`	ADI	MED	CISI
F-Combsum	39.14	28.23	33.63
F-Combmax	37.97	23.21	32.45
F-Combmnz	41.31	19.11	37.68

The null hypothesis is rejected successfully and it is proved that the filter size has impact over the combination function. Obviously the filter size which has higher average value will become the optimal filter size. Before computing the optimal filter size the scores are normalized to avoid the domination of the data set which has the higher relevance score range. The average values of the normalized score for the three functions over the data sets are given in the Table 5.

Table 5. Average precision value at various filter size

	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
ADI	0.87	0.89	0.89	0.97	0.98	0.99	1.0	0.99	0.99	0.99
MED	0.91	0.96	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.97
CISI	0.88	0.92	0.95	0.95	0.95	0.99	0.99	0.99	0.99	0.99

The graph (Figure 3) shows the overall average value and it eases the comparison process.

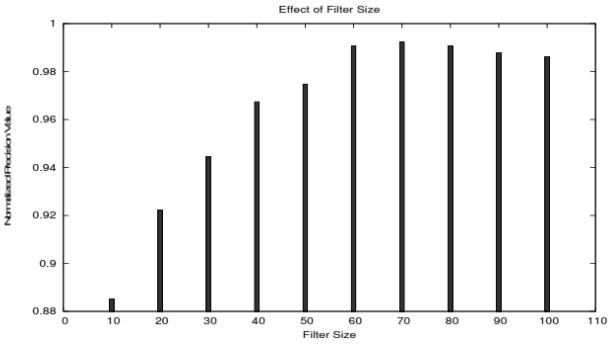

Fig. 3. Variations in performance at various filter size

As the peak precision value occurs at a point X corresponds to the 70% of the filter size at the flattening point, OX is fixed as the optimal filter size. Figure 4 shows the algorithm to determine the optimal size and the method of assigning the relevance score to the documents.

1. Calculate the absolute value of maximum difference among all relevance scores (d).

2. Let the size of the filter at the flattening point as $d_1 = 1 - d$.

3. Convert it in to decibel equivalent as d1 (in dB) = $20 * \log(d_1)$.

4. Calculate the optimal filter size as $0.7 * d_1$.

5. Let the overlap value (the number of relevant scores that lie inside the filter) be γ

6. Apply the Calculated overlap value to the CombFunctions to derive the F-CombFunctions

F-CombMAX - Maximum of all relevance sores $\times\gamma$

F-CombMNZ - Sum of all relevance scores $\times\gamma$

F-CombSUM - Sum of all relevance scores lie inside the filter

Fig. 4. Algorithm for calculating the relevance score using F-CombFunctions

6. Experiment and results

The benchmark test collections and the retrieval schemes mentioned in §4.1 given by (7) - (11) are used to test the effectiveness of the proposed functions. The 11- point interpolated precision measure is used for comparing the performance of the newly defined filter based fusion functions with the conventional CombFunctions.

6.1 Number of schemes to be fused

In the experiment, a total of seven retrieval schemes are used ((7) - (11)) and it is planned to test the performance of various combinations of them. Hence, varying number of schemes starts from '2' and ends with '7' are used in the experiment. There are possible 21, 35, 35, 21, 7 and 1 combinations are available for the 2, 3, 4, 5, 6, and 7 number of schemes respectively (7C2, 7C3 , 7C4 , 7C5 , 7C6 , 7C7). Average of 11-pt interpolated precision of all combinations is recorded for comparison purpose.

6.2 Results

Table 6 shows the average 11-pt interpolated precision of the F-CombFunctions and CombFunctions.

Table 6. Precision values for F-Comb and CombFunctions

No of Schemes	Comb SUM	F–Comb SUM	Comb MNZ	F–Comb MNZ	Comb MAX	F–Comb MAX
ADI						
2	0.3459	0.3495	0.3459	0.3484	0.3448	0.3505
3	0.3481	0.3520	0.3481	0.3510	0.3438	0.3511
4	0.3462	0.3509	0.3462	0.3499	0.3435	0.3537
5	0.3463	0.3516	0.3463	0.3508	0.3441	0.3574
6	0.3434	0.3507	0.3434	0.3501	0.3455	0.3616
7	0.3413	0.3477	0.3413	0.3467	0.3475	0.3623
CISI						
2	0.1851	0.1886	0.1851	0.1888	0.1853	0.1898
3	0.1879	0.1915	0.1879	0.1920	0.1849	0.1902
4	0.1897	0.1934	0.1897	0.1942	0.1856	0.1912
5	0.1905	0.1941	0.1905	0.1950	0.1869	0.1926
6	0.1908	0.1943	0.1908	0.1954	0.1885	0.1940
7	0.1911	0.1914	0.1911	0.1925	0.1901	0.1925
MED						
2	0.4825	0.4855	0.4825	0.4829	0.4521	0.4619
3	0.4983	0.5015	0.4983	0.5020	0.4539	0.4706
4	0.5049	0.5083	0.5048	0.5091	0.4548	0.4685
5	0.5085	0.5118	0.5085	0.5127	0.4551	0.4675
6	0.5104	0.5136	0.5104	0.5146	0.4548	0.4701
7	0.5143	0.5146	0.5143	0.5160	0.4541	0.4633

The graph (Figure 5) shows the 11-point interpolated precision for all functions over the three test data sets.

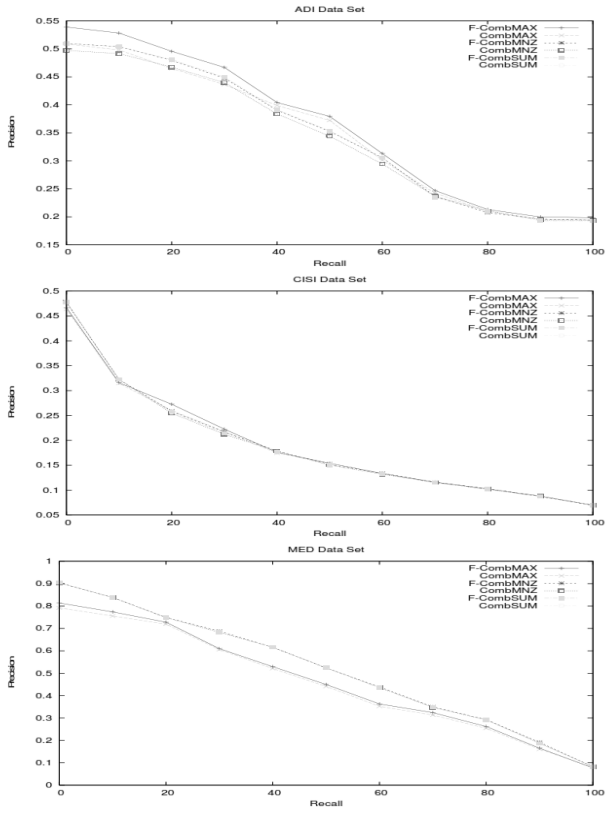

Fig. 5. 11-Point interpolated precision value

6.3 Performance comparison and test of hypothesis

Performance of F-CombFunctions is compared with the CombFunctions and the percentage of improvement for F-CombFunctions is computed. Paired 'student-t' test is also used for comparison purpose. In this test μ_1 represent the average precision value for CombFunctions and μ_2 represents the average precision value for F-CombFunctions. The null and alternative hypotheses are shown below.

$$H0: \mu 1 = \mu 2$$

$$H1: \mu 1 \leq \mu 2$$

The Table 7 gives the percentage of improvement and 't' value for the F- CombFunctions. Further, it can be seen that there is no t value for the combination function that merges 7 retrieval schemes.

Table 7. % of Improvement and `t' values for F-CombFunctions

Function	2		3		4		5		6		7
	%	t	%	t	%	t	%	t	%	t	%
ADI											
sum	1.07	4.25	1.13	7.10	1.36	9.48	1.54	8.29	1.24	5.18	1.87
mnz	0.72	2.29	0.84	4.13	1.06	5.76	1.29	6.74	1.93	5.05	1.60
max	1.65	3.33	2.11	5.73	2.98	8.19	3.89	8.39	4.68	8.96	4.24
CISI											
sum	1.90	7.41	1.92	9.83	1.92	8.72	1.86	6.74	1.82	5.25	0.13
mnz	1.99	6.85	2.20	4.69	2.37	7.39	2.35	8.86	2.39	6.43	0.73
max	2.42	5.60	2.86	7.36	3.00	5.85	3.03	5.85	2.89	4.95	1.25
MED											
sum	0.63	4.78	0.65	8.84	0.69	7.59	0.64	6.54	0.63	5.73	0.05
mnz	0.07	0.22	0.75	5.88	0.84	6.32	0.83	5.23	0.81	4.66	0.33
max	2.16	4.78	3.67	4.95	3.02	5.17	2.74	4.33	3.37	6.67	2.02

The column F-CombMAX in the Table 4 indicates that the improvement in performance for the F-CombMAX function is significantly higher as it utilizes the advantage of both skimming and Chorus effects at the optimal filter.

7. Conclusion

The statistical communication theory indicates that the deletion of low relevance scores improves the performance of the fusion functions. Effect of filter size on fusion function is analyzed and the results are used to find out the optimal filter size. The performance of the fusion functions within the optimal filter is found to be better as all ill-performing schemes are deleted. In this work, we propose new fusion functions namely F-functions. The F-CombMAX achieves significant improvement over the others and hence it may be advantageously used for IR. We achieved a maximum of 13.2%, and an average of 3.69% performance improvement. This work is very useful for selecting the search engines in meta search engine. The current meta-search engines use the static method for search engine's selection. As our proposed method consider all search engines equally and select them dynamically, it will improve the meta-search engine's performance. In near future, we want to test the F-functions over the web search engines.

References

Belkin, N. Kantor, P. & Quatrain, R. (1994). Combining evidence for information retrieval. Proceedings of the 2nd Text REtieval Conference', Gaithersburg, Maryland, USA: 35–44.

Bilhart, H. (2003). Learning retrieval expert combinations with genetic algorithm. International Journal of Uncertainty, Fuzziness and Knowledge- Based Systems, **11**(1):87–114.

Croft, W.B. (2002). Combining approaches to information retrieval. Advances in information retrieval: 1–36.

Fisher, H. L. & Elchesen, D. R. (1972). Effectiveness of combining title words and index terms in machine retrieval searches. Nature, **238**(5359):109–110.

Fox, E.A. & Shaw, J.A. (1994). Combination of multiple searches. Proceedings of the Second Text Retrieval Conference: 243–252.

Fox, E.A & Shaw, J.A. (1995). Combination of multiple searches. Proceedings of the Third Text Retrieval Conference: 105–108.

Ghosh, K. Parui, S. K. & Majumder, P. (2015). Learning combination weights in data fusion using genetic algorithms. Information Processing & Management, **51**(3): 306–328.

Korfhage, R.R. (1997). Information storage and retrieval. Willey Computer Publishing.

Lee, J.H. (1995). Combining multiple evidence from different properties of weighting schemes. Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Wasington, USA: 180–188.

Lee, J.H. (1997a). Analyses of multiple evidence combination. Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, Philadelphia, PA, USA: 260–276.

Lee, J.H. (1997b). Combining multiple evidence from different relevant feedback networks. Proceedings of the 5th international Conference on Database Systems for Advanced Applications, Melbourne, Australia: 421–430.

Ponte, J.M. & Croft, W.B. (1998). A language modeling approach to information retrieval. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM: 275–281.

Salton, G. & McGill, M. (1986). Introduction to modern information retrieval. McGraw–Gill.

Sanke, M. (2015). A technique of data fusion for effective text retrieval. International Journal of Computer Applications, **111**(8):5-9.

Shannon, C.E. (1948). A mathematical theory of communication. The Bell System Technical Journal **27**:379–423 and 626–656.

Ullah, A., Baharudin, B. *et al.* (2016). Pattern and semantic analysis to improve unsupervised techniques for opinion target identification, Kuwait Journal of Science, **43**(1):129-149.

Vogt, C.C. (1999). Adaptive combination of evidence for information Retrieval, PhD thesis, University of California, Sandiego.

Wu, S. (2012). Data fusion in information retrieval. Springer Science & Business Media, Springer - Verlag Berlin Heidelberg Ed. Pp 228.

Wu, S. Li, J. Zeng, X. & Bi, Y. (2014). Adaptive data fusion methods in information retrieval. Journal of the Association for Information Science and Technology, **65**(10):2048–2061.

Yates, R.B. & Neto, B.R. (1999). Modern information retrieval. Pearson Education, Pp 240.

Submitted **:** 09/09/2015
Revised **:** 23/04/2016
Accepted **:** 25/04/2016

اختيار فعال لأنظمة الاسترجاع لدمج البيانات

كريشنان باتري*[1]
قسم هندسة الالكترونيات والاتصالات
بسنا PSNA كلية الهندسة والتكنولوجيا، ديندیجول، الهند
Krishnan.batri@gmail.com

خلاصة

دمج النتائج من عدد أكبر من أنظمة / برامج استرجاع قد يعزز من أداء نظام استرجاع المعلومات. يكمن نجاح الدمج في الأنظمة الأعضاء المُختارة. يستكشف هذا البحث خوارزمية اختيار فعالة مشتقة من مفهوم الفلترة (the filter concept)، من خلال معاملة أنظمة الاسترجاع منخفضة الدرجة مثل الضوضاء. تم اختبار الخوارزمية المقترحة على مجموعات اختبار قياس الأداء المعيارية الثلاثة، وهي: معهد التوثيق الأمريكي (ADI)، مركز الاختراعات والمعلومات العلمية (CISI)، وميدلارز[i](Medlars) (MED). تم اختبار تطابق النتيجة المحوسبة بواسطة اختبار تي لعينة الازواج (paired student t-test[ii]). لوحظ أن الخوارزمية المُقدمة أسفرت عن تحسن ملحوظ بالمقارنة مع دوال الدمج الحالية. يرجع تحسن الأداء في النهج المُقترح إلى الانخفاض في تأثير جوقة التضخم (amplification chorus) الناجم عن أنظمة استرجاع منخفضة الدرجة.

i Medlars: Medical Literature Analysis and retrieval system تحليل النشرات الطبية واسترجاعها

ii paired student t-test: A paired t-test is used to compare two population means where you have two samples in which observations in one sample can be paired with observations in the other sample.