Effect of Weight Assignment in Data Fusion Based Information Retrieval

Krishnan Batri[1], Veerasamy Murugesh[2] and Nagammapudur Gopalan[3]
[1]Department of Computer Engineering, European University of Lefke, Turkey
[2]Department of Microsoft Information Technology, Keimyung University, Republic of Korea
[3]Department of Computer Applications, National Institute of Technology, India

Abstract: *Variation in performances of an Information Retrieval system, which merges results from a number of retrieval schemes possessing equal and unequal weights, is studied in this paper. Weight of the retrieval schemes for a particular document is derived from the relevance scores of that corresponding document. Since, the relevance scores are varying from document to document and corpus to corpus, the method proposed is dynamic. A number of weight calculation methods, which are using the error value for computation purpose, are discussed in this paper. The effectiveness of the weight calculation is tested over three benchmark test collections viz., ADI, CISI and MED. It has been identified that the methods discussed in this paper retrieve articles effectively and they are independent of history or any training data.*

Keywords: *Data fusion, information retrieval, relevance score, similarity measure, and weight assignment.*

Received September 16, 2008; accepted August 3, 2009

1. Introduction

Repository of digital data is booming day-by-day [13] and finding the needy information becomes a tedious task. The process associated with the searching and selection of relevant articles from a collection (corpus) is termed as Information Retrieval (IR) [10]. A number of IR models and schemes [4] are proposed to ease the retrieving task. The retrieval schemes select the pertinent documents from the corpus based on the relevance scores, which are obtained from the similarity measures [13]. Similarity measures assign relevance score to documents based on the match between the documents and the users' given query. Effectiveness of the various retrieval schemes is tested by using Precision and Recall measures [9] and it has been identified that the retrieval schemes have the drawback of inconsistent performance [14].

Fusion is the process of combining data from multiple sources [5, 15]. In IR, the fusion may merge the following [11]: results from the multiple document representations (representation fusion), multiple query forms (query fusion), multiple systems (system fusion) and multiple retrieval schemes (method fusion). Early literatures [13] indicate that, the fusion consistently yields better results by overcoming the drawback i.e., inconsistent performance.

This paper addresses the effect of weight assignment for the retrieval schemes along with the effective methods for assigning the weights. Two possible ways are available for the weight assignments viz., Adaptive and Non-Adaptive. In adaptive methods, weights are learned from history, training data and user feedbacks. In non-adaptive methods, weights are derived from prior knowledge about the retrieval systems. This article deals with the newly proposed non-adaptive weight assignment methods. The normalized relevance scores of the documents are used for weight calculation. As the scores of the documents are varying, the assigned weights will vary; hence method proposed is dynamic. Performance of the methods is tested over three benchmark test collections namely: ADI, CISI and MED. `Paired Student-t' test is used to compare the effectiveness of the linear combination method and the proposed weight assignment based merging. The experimental results and the computed `t' values look promising.

The rest of the paper is organized as follows. The detailed discussion about data fusion and the three effects involved in the fusion are presented in next section. The prior works in the area of data fusion are given in section 3. The section 4 discusses the information content analysis of the retrieval schemes and section 5 proffers the discussion about various weight calculation methods. The section 6 shows the experimental results and section 7 concludes with the future direction of research.

2. Data Fusion

Combining or merging the results from a number of sources is termed as fusion. By doing so, it effectively taps the merits of all participating members. In IR, the fusion function, which assigns the final relevance score to a document based on the returned relevance scores uses the following effects [11] viz., :

1. Skimming effect.
2. Chorus effect.
3. Dark horse effect.

- Skimming effect: the retrieval schemes used to collect relevant documents and 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 them are selected, then the phenomenon is termed as skimming effect.
- Chorus effect: the chorus effect assigns a high degree of relevance to the documents found in a majority of lists returned by the retrieval schemes.
- Dark horse effect: the dark horse is one in which the documents may get unusually accurate (or inaccurate) relevance score.

It has been observed that the functions based on chorus effect yield better results [5, 6] and outperform the others based on either skimming or dark horse effects.

2.1. Chorus Effect

Two heads are better than one' is the basic notion of chorus effect. The fusion functions, which are based on chorus effect, declared a document as a relevant one, if more number of retrieval schemes suggest that particular document as relevant.

During fusion, the chorus effect may get amplified by few retrieval schemes, whose scores are significantly differ form others. This perturbs the fusion by creating an illusion about the relevance of those documents. Controlling the disturbances caused by ill performing schemes may reduce the amplification of chorus effect.

The proposed weight assignment methods assign low weights to the worst performing schemes. The low weights diminish the contribution of the respective scheme and successfully negotiated the disturbances caused by them. As a result, the amplification of chorus effect gets reduced and it leads to performance improvement.

3. Fusion Techniques

The advantages of fusion were explored by Fisher [1] in early 70's. He successfully merged two Boolean searches operating on the title words and manually generated index terms. The results of Fisher's experiment show improvement in performance for fusion against the best search.

The Comb-functions, which was introduced by Fax and shaw [2, 3], merge more number of schemes in comparison with the Fisher's method. The Table 1 shows the various Comb-functions.

Table 1. Comb-function for combining scores.

Function Name	Explanation
CombMIN	Minimum of Individual Relevance Scores
CombMAX	Maximum of Individual Relevance Scores
CombSUM	Sum of Individual Relevance Scores
CombANZ	CombSUM ÷ Number of Nonzero Relevance Scores
CombMNZ	CombSUM × Number of Nonzero Relevance Scores

Lee [5, 6, and 7] further explored the Comb-functions and proposed some new rationales and indicators for fusion. The weighted linear combination [12, 16], which is the successor of linear combination method [11], combines more number of schemes. The linear combination method (for example CombSUM) sums up all scores. The final relevance score `r' of a document `d' assigned by the weighted linear combination for a query `q' is given in equation 1.

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

where, r_j - final relevance score of the j^{th} document, θ_i - weight of the i^{th} retrieval scheme [θ value is user defined, $-\infty < \theta < +\infty$], R_i - relevance score returned by the i^{th} retrieval scheme for the j^{th} document and k - number of retrieval schemes to be fused.

The effectiveness of weight assignment schemes is tested by comparing its performance against the linear combination method.

The weighted linear combination method has the limitation of requiring prior knowledge, history and training data for weight assignment. There are various learning tools are available that include Genetic algorithm, Neural network etc., the statistical concept like regression may also be used for weight calculation [12]. These learning or predicting methods use training data and the computed weights are remain fixed until the next learning or predicting process. Hence, the weights in these methods are static.

The proposed weight assignment methods compute the weights from the relevance scores. Hence, it is independent of history and training data.

4. Information Content

The scores assigned by the retrieval schemes give information about the relevance of the documents. Hence, the certainty about the significance of the documents may be analyzed by using the statistical information theory [11]. As the retrieval schemes give information about the relevance of a document, they may be treated as the message symbols for further analysis and the fusion functions which operate on all retrieval schemes may be treated as typical message sources.

Let `S' be the message symbol set and `P' is the set of probability (p) of their occurrence. If there are `n' retrieval schemes then S and P become:

$S = \{s_1, s_2, \ldots\ldots, s_n\}$

$$P = \{p_1, p_2, \ldots\ldots, p_n\} \text{ with } \sum_{i=1}^{n} p_i = 1 \tag{2}$$

Let R_j be the relevance score returned by the j^{th} retrieval scheme for a particular document then the probability (p_j) of message symbol s_j for that document may be calculated as

$$p_j = \frac{R_j}{\sum_{i=1}^{n} R_i} (1 \le j \le n) \tag{3}$$

where, n - Number of retrieval schemes participating in fusion. Hence, it may be inferred that

$$p_j \propto R_j$$
$$p_j \propto \frac{1}{\sum_{i=1}^{n} R_i} \tag{4}$$

As the intention of this analysis is to improve the performance of the fusion function, the effectiveness of the message source should be analyzed. Entropy is used for this purpose and the formula used for this calculation is given by:

$$H(Y) = -\sum_{i=1}^{n} p_i . \log_{10} p_i \tag{5}$$

The entropy will reach its maximum only when all probabilities are equal. As the probabilities are computed from the relevance scores, the scores should be equal. Consider a scenario, where the scores of a document are equal. Under this condition:

$$p_j = \frac{R_j}{\sum_{i=1}^{n} R_i} = \frac{R_j}{n.R_j} = \frac{1}{n} \tag{6}$$
$$p_j = \frac{1}{n}$$

Now compute the entropy for this using the newly calculated probability value.

$$H(Y) = \log(n) \tag{7}$$

The fusion function yields its maximum capacity only when all sources are equal. In view of the statistical information theory, the desired criteria for fusion may be stated as `` the relevance scores of a document should be equal".

In a real time scenario, all scores of a document becoming equal are impossible. Hence, it is necessary to delete or reduce the contribution of retrieval schemes whose scores are significantly different from others. In the second case, low weights are assigned to the ill performing schemes to minimize its contribution.

5. Weight Assignment Method

The weight assignment method modifies the contribution of the retrieval schemes by assigning weights to them. The weights for the schemes are derived from their performance. The contribution of the best performing schemes are boosted by assigning higher weights to them and the ill performing schemes are suppressed by assigning low weights to them. The formula used to calculate the probability based on the weighted relevance score may be restated as:

$$p_i = \frac{w_i . R_i}{\sum_{j=1}^{n} w_j . R_j} \tag{8}$$

where, w_i -- weight of the i^{th} retrieval scheme for a particular document. The assigned weights successfully modify the contribution of the retrieval schemes and it is explained with an example given in Table 2. As a result of the weight modification, the value of the entropy gets changed.

Table 2. Change in entropy value: an example.

S.No	Score	Weight	Modified Score	Probability
1	0.1	1	0.1	0.110351
2	0.11	0.999989	0.109999	0.121385
3	0.115	0.999974	0.114997	0.126901
4	0.2	0.998	0.1996	0.220261
5	0.9	0.424	0.3816	0.421101
Entropy Without Weights				0.50079
Entropy With Weights				0.63347

5.1. Weight Calculation

The weight calculation used in this chapter is adapted from perceptron learning principle [8]. The formula used for weight calculation in perceptron learning is given below.

$$w^t = w^{t-1} + c.(e).x_t \tag{9}$$

where, w^t - Weight for the current iteration, w^{t-1}- weight of the previous iteration, c - Learning constant, e - Difference between the desired and actual values (d-a) and, x_t - Current input value (actual value) in perceptron learning, value of w^t depends on w^{t-1}. From 9, it is inferred that the weights are either incremented or decremented, when the actual value is less or greater than the desired value.

Consider the adaptation of weight calculation formula for data fusion. As, the equal relevance scores become the desired criteria for fusion, the higher or lower actual values (a) in comparison with the desired value (d) will disturb the performance. So, weights should be decremented in accordance with the error values and the sign `+' in equation 9 is to be replaced with `-'. If, the weights of the retrieval scheme for a document are derived from the preceding iterations, then it leads to degradation in effectiveness. Hence, the term w^{t-1} is replaced with the constant a `k'

(Experiments are conducted at various k values and based on the results the k is fixed at `5'). As there is no learning process, the learning constant is deleted from equation 9. The formula used for calculating the weights based on the proposed method is given in equation 8.

$$w_{ij} = k - (e).R_{ij} \tag{10}$$

where, w_{ij} - weight of the retrieval scheme `i' for the document `j', k - constant (value of `k' is fixed as `5' based on the trial and error) and R_{ij} - relevance score assigned to the document `j' by retrieval scheme `i'.

The entire weight assignment process depends on the error value (e), which in turn depends on desired (d) and actual (a) values. The desired value is an imaginary relevance score for which the fusion becomes less involved. The actual value is the relevance score assigned to the documents by the retrieval schemes. Under this circumstance, the error value may be calculated as:

$$e = d - a \tag{11}$$

Care should be taken to select the desired value. In this paper, five desired values are used. They are:

a. One.
b. Max.
c. Zero.
d. Min.
e. Average.

Discussion on the above is presented in section 5.1.1. The error values based on the `Average' desired value may posses both positive and negative numbers. The negative value, which increments the weights beyond the value of `k', leads to performance degradation. Hence, the squared error value is used and the final formula used for weight calculation is given by:

$$w_{ij} = k - e^2.R_{ij} \tag{12}$$

5.1.1. Desired Values

The retrieval schemes, which assign relatively higher or lower scores, perturb the fusion. So, the selected desired value should suppress both undue higher and lower scores. For this purpose the following five desired values are selected.

- One: in the experiment, the scores returned by the schemes are normalized and are in range of 0 - 1. Hence, the desired value may be select as one.
- Max: in practical situations, the chance for the relevance score being `1' is rare. Hence, the maximum of scores for a particular document returned by all retrieval schemes is treated as the desired value of that particular document.
- Reason for choosing one and max: the desired values `Max' and `One' boost the contribution of higher magnitude relevance scores. It assumes that higher relevance scores aid the performance and lower scores amplify the chorus effect. Both the `Max' and `One' desired values give low weights to the schemes which assign low scores.
- Zero: since, the lower edge of the normalized relevance score is `Zero', the desired value may be selected as `0'.
- Min: instead of considering `0' as minimal, the minimum of scores returned by all the schemes is considered.
- Reason for Choosing Zero and Min: the desired value of `Zero' and `Min' boost the contribution of the low relevance scores and suppress the others. The desired values of zero and min assume that, the chorus effect gets amplified by some retrieval schemes, which assign relatively higher values. The amplification may be reduced by assigning low weights and it may increase their performance.
- Average: The average of all relevance scores is treated as the desired value.

6. Experiment and Results

Experiments are conducted to test the impact of the proposed weight assignment methods. For the comparison purpose, average 11-point interpolated precision measure is used.

6.1. Experimental Setup

6.1.1. Datasets

The experiments are conducted on three bench mark test document collections viz.,:

1. ADI.
2. CISI.
3. MED under a uniform environment.

Table 3 shows the characteristics of the three datasets.

Table 3. 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

6.1.2. Index Term Processing

The unwanted words from the corpus are removed by using the stop-word list (Smart stop word list) and the remaining words are trimmed by the help of stemmer algorithm (Porter stemmer). Formulas used to assign weights to index terms are given in (13) and (14). Term Frequency-Inverse Document Frequency (TF-

IDF) weight assignment method is used for this purpose.

$$w_t = \log_{10}(1+\frac{N}{f_t}) \quad (13)$$

$$w_{d,t} = r_{d,t}.w_t \quad (14)$$

where, N - total number of document in the corpus, f_t - number of document 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.

6.1.3. Retrieval Schemes

Two different retrieval models are used in the experiment to minimize the domination of any particular model. The similarity measures of the Vector Space Model (VSM) and `p-norm model' [4] are used as the retrieval schemes in the experiment. Similarity measures of VSM used in the experiment are given in equations (15-18).

$$\text{Cosine Similarity} \quad R(q,d) = \frac{\sum_{i \in q \cap d} w_{q,t}.w_{d,t}}{W_q.W_d} \quad (15)$$

$$\text{Inner Product} \quad R(q,d) = \sum_{i \in q \cap d} w_{q,t}.w_{d,t} \quad (16)$$

$$\text{Dice Coefficien t} \quad R(q,d) = \frac{2.\sum_{i \in q \cap d} w_{q,t}.w_{d,t}}{W_q^2 + W_d^2} \quad (17)$$

$$\text{Jaccard} \quad R(q,d) = \frac{\sum_{i \in q \cap d} w_{q,t}.w_{d,t}}{W_q^2 + W_d^2 - \sum_{i \in q \cap d} w_{q,t}.w_{d,t}} \quad (18)$$

where, R - relevance 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 is also used as a retrieval scheme in the experiment and it is shown in equation 19.

$$R(q_{and}, d_j) = 1 - \left(\frac{(1-w_1)^p + (1-w_2)^p + ...+(1-w_m)^p}{m} \right)^{1/p} \quad (19)$$

where, w_m - weight of the m^{th} index term and $1 \le p \le \infty$. The p value is set as 1.5, 2.5 and 3.5 based on trial and error method. The above seven retrieval schemes (VSM 4 and p-norm 3) are used to test the effectiveness of the proposed weight assignment method.

6.1.4. Normalization

The scores returned by the various retrieval schemes based on weight of the index terms are of various ranges. The scheme, which posses higher range, dominates the fusion. In order to maintain a uniform environment, normalization is used and in the experiment `Max normalization' is selected. The formula used for `Max normalization' is given in equation 20.

$$R_{normalized} = \frac{R_{unnormalized}}{R_{\max}} \quad (20)$$

where, $R_{unnormalized}$ - relevance score returned by a retrieval scheme and R_{max} - Maximum relevance score returned by a generic retrieval scheme.

6.1.5. Number of Schemes to be Fused

In the experiment, a total of seven retrieval schemes are used and the performance of their various combinations ($7C_i$, i=2,3,...,7) are tested. Average of 11-pt interpolated precision of all combinations is recorded for comparison purpose.

6.2. Results

6.2.1. Results of CombSUM

CombSUM function linearly combines the score and the performance of the proposed weight assignment method is compared with the CombSUM function. Table 4 gives the average 11-point interpolated precision value for CombSUM function.

Table 4. Avg 11-pt interpolated precision for combSUM.

No. of Schemes	ADI	CISI	MED
2	0.3459	0.1851	0.4825
3	0.3481	0.1879	0.4983
4	0.3462	0.1897	0.5048
5	0.3463	0.1905	0.5085
6	0.3434	0.1908	0.5104
7	0.3413	0.1911	0.5143

6.2.2. Results of the Weight Assignment Method

The Table 5 gives the results of weight assignment method for various desired values and varying number of retrieval schemes.

Table 5 Avg 11-pt Interpolated precision for weight assignment method.

Des Value	2	3	4	5	6	7
ADI						
Zero	0.3539	0.3568	0.3587	0.3594	0.3613	0.3556
One	0.3520	0.3546	0.3524	0.3528	0.3420	0.3432
Ave	0.3518	0.3546	0.3529	0.3537	0.3563	0.3475
Min	0.3519	0.3556	0.3563	0.3587	0.3583	0.3528
Max	0.3520	0.3547	0.3529	0.3532	0.3503	0.3475
CISI						
Zero	0.1935	0.1967	0.1985	0.1987	0.1988	0.1985
One	0.1911	0.1941	0.1958	0.1966	0.1972	0.1977
Ave	0.1912	0.1942	0.1960	0.1967	0.1971	0.1974
Min	0,1912	0.1943	0.1961	0.1972	0.1975	0.1982
Max	0.1912	0.1941	0.1959	0.1968	0.1972	0.1975
MED						
Zero	0.4917	0.5092	0.5167	0.5208	0.5241	0.5267
One	0.4887	0.5043	0.5110	0.5146	0.5160	0.5197
Ave	0.4908	0.5061	0.5128	0.5173	0.5197	0.5230
Min	0.4913	0.5085	0.5161	0.5207	0.5236	0.5271
Max	0.4886	0.5046	0.5120	0.5149	0.5166	0.5200

6.2.3. Performance Comparison

The various weight assignment methods are compared with CombSUM function to evaluate it's performance. Percentage of improvement is calculated to make the comparison process an easier one. Paired `t' test is used to compare the performance based on the hypothesis. In the test, μ_1 represents the average precision value of CombSUM and μ_2 represents the average value of the weight assignment method. The null and alternative hypotheses are shown bellow. Null Hypothesis: H_0: $\mu_1 = \mu_2$, Alternative hypothesis: H_1: $\mu_1 < \mu_2$.

The Table 6 shows the percentage of improvement and `t' value for various weight assignment method against the CombSUM and column `7' has no `t' value because there is only one combination. From the table it is identified that desired value of `Zero' outperforms the other weight assignment methods.

Table 6. `t' Value and % of Improvement for weight assignment method.

Des Value	2		3		4		5		6		7
	%	t	%	t	%	t	%	t	%	t	%
ADI											
Zero	2.3	3.9	2.5	3.1	3.6	7.1	3.8	5.2	5.2	5.8	4.2
One	1.8	1.9	1.9	3.1	1.8	2.9	1.9	1.7	1.8	3.8	0.6
Ave	1.7	0.3	1.9	2.6	1.9	2.0	2.1	3.7	3.7	2.5	1.8
Min	1.8	2.4	2.2	3.8	2.9	5.5	3.6	4.5	4.3	4.7	3.4
Max	1.8	2.7	1.9	3.6	1.9	4.4	2.0	4.1	2.0	2.9	1.8
CISI											
Zero	4.6	4.0	4.7	6.0	4.6	5.8	4.3	3.9	4.2	2.1	3.9
One	3.2	2.6	3.3	3.2	3.2	3.5	3.2	2.9	3.3	2.1	3.4
Ave	3.3	4.5	3.4	4.5	3.3	4.1	3.3	3.6	3.3	1.5	3.3
Min	3.3	4.0	3.4	2.9	3.4	4.0	3.5	4.5	3.5	4.2	3.7
Max	3.3	2.7	3.3	3.2	3.3	3.2	3.3	2.4	3.4	2.5	3.2
MED											
Zero	1.9	7.2	2.2	8.2	2.4	9.9	2.4	9.4	2.7	4.4	2.4
One	1.3	1.9	1.2	0.8	1.2	1.3	1.2	1.2	1.1	1.8	1.1
Ave	1.7	5.0	1.6	5.8	1.6	8.6	1.7	5.5	1.8	5.2	1.7
Min	1.8	5.2	2.1	8.0	2.2	9.8	2.4	8.8	2.6	3.7	2.5
Max	1.3	0.9	1.3	2.2	1.2	1.3	1.3	2.8	1.2	0.5	1.1

The weight assignment methods boost the contribution of the retrieval schemes whose scores are in range with others. The `Zero' and `Min' methods suppress the higher relevance scores and the `Max' and `One' performing the opposite work. Table 7 shows the overall average precision value in descending order.

Table 7. Overall average precision value in descending order.

Desired Value	Precision Value
Zero	0.356643
One	0.355293
Ave	0.35329
Min	0.352169
Max	0.351743

From the table it has been identified that the disturbances caused by the schemes, which are returning low relevance score, is minimal in comparison with others. As zero and min methods effectively suppress the higher relevance score they yield better performance.

7. Conclusions

In this paper, various weight assignment methods are proposed. Out of five methods, the desired value `Zero' performs consistently well. This is due to the fact that, the disturbances caused by the amplification of chorus effect produced by low range schemes are minimal in comparison with it's counter parts. In future, it is planned to evolve a new method that combines the merits of all weight assignment methods discussed in this paper.

References

[1] Fisher H. and Elchesen D., "Effectiveness of Combining Title Words and Index Terms in Machine Retrieval Searches," *Computer Journal of Nature*, vol. 238, no. 5, pp. 109-110, 1972.

[2] Fox E. and Shaw J., "Combination of Multiple Searches," *Second Text Retrieval Conference*, pp. 243-252, 1994.

[3] Fox E. and Shaw J., "Combination of Multiple Searches," *Third Text Retrieval Conference*, pp. 105-108, 1995.

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

[5] Lee J., "Combining Multiple Evidence from Different Properties of Weighting Schemes," *in Proceedings of 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 180-188, 1995.

[6] Lee J., "Analyses of Multiple Evidence Combination," *in Proceedings of 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 260-276, 1997.

[7] Lee J., "Combining Multiple Evidence from Different Relevant Feedback Networks," *in Proceedings of 5th International Conference on Database Systems for Advanced Applications*, pp. 421-430, 1997.

[8] Luger G., *Artificial Intelligence*, Pearson Education, 2002.

[9] Nassar M. and Kanaan G., "The Factors Affecting the Performance of Data Fusion Algorithms," *in Proceedings of ICIME*, pp. 465-470, 2009

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

[11] Shannon C., "A Mathematical Theory of Communication," *Computer Journal of the Bell*

System Technical, vol. 27, no. 3, pp. 379-423, 1948.
[12] Shengli W., "Applying Statistical Principles to Data Fusion in Information Retrieval," *Expert Systems with Applications: An International Journal*, vol. 36, no. 5, pp. 2997-3006, 2009.
[13] Vogt C., "How Much More is Better? Characterizing the Effects of Adding More Ir Systems to a Combination," *in Proceedings of Content-Based Multimedia Information Access*, pp. 457-475, 2000.
[14] Vogt C. and Cottrell G., "Fusion via a Linear Combination of Scores," *Computer Journal of Information Retrieval*, vol. 3, no. 1, pp. 151-173, 1999.
[15] Yates R. and Neto B., Modern Information Retrieval, *Pearson Education*, 1999.
[16] Zobel J. and Moffat A., "Exploring the Similarity Space," *Computer Journal of ACM SIGIR Forum*, vol. 32, no. 1, pp. 18-34, 1998.

Krishnan Batri received the Master of engineering from Madurai Kamarj University in 2003. Completed his PhD in computer science and engineering from National Institute of Technology, Tiruchirappalli, Tamilnadu, India. Currently he is working as a assistant professor with the department of computer engineering at European University of Lefke, TRNC, Mersin-10, Turkey. His research interests include information retrieval, data fusion and genetic algorithms.

Verasamy Murugesh obtained his Bachelor of Science in Computer Science and Master of Computer Applications from Bharathiar University, Coimbatore, India during 1992 and 1995 respectively. He completed his PhD in Computer Science from Bharathidasan University, Tiruchirappalli, India during the year 2006. He has held various positions at National Institute of Technology, Tiruchirappalli, India, Hannam University, Republic of Korea and European University of Lefke, TRNC Mersin 10, Turkey. Currently, he is working as Associate Professor in the Department of Microsoft Information Technology, Keimyung Adams College, Keimyung University, Republic of Korea. His fields of interest are in Neural Network based Image Processing and Scientific computing. He has published more than 40 papers in international, national journals and conferences.

Nagammapudur Gopalan received the M.Sc from Madras University in 1978 and the PhD in applied mathematics from Indian Institute of Science, Bangalore, in 1983. Currently he is a Professor with 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.