International Journal of Soft Computing 2 (1): 11-16, 2007

Adaptive Selection of Top-m Retrieval Strategies for Data Fusion in Information Retrieval

N.P. Gopalan and K. Batri
Department of Computer Applications, NIT, Tiruchirappalli-15
Department of Computer Science and Engineering, NIT, Tiruchirappalli-15

Abstract: Data fusion for Information Retrieval (IR) usually combines various strategies (schemes) to enhance the performance. The process of selecting the top `m' strategies with appropriate weights has been presented in this study. Genetic Algorithm (GA) with relevant judgments obtained for the user specified queries as training data is employed for the adaptive selection of the suitable strategies. The significance of experimental results obtained with three benchmark test collections is examined using Student-t test. The present adaptive method has been observed to perform consistently well in comparison with the individual ones participating in the fusion.

Key words: Information retrieval, data fusion, adaptive selection, genetic algorithm, student-t test

INTRODUCTION

The process of finding the needy information from it's repository is a non-trivial task[1] and it is necessary to formulate a process that effectively proffers the pertinent documents. The task of retrieving germane article is termed as Information Retrieval (IR)[2]. Various models, the loci-classics of IR and strategies have been proposed to represent and organize the document collection with less effort.

A 'Model' is a set of premises[3] with an algorithm for ranking documents with regard to a user query. A 'Strategy' or scheme is a method of assigning similarity score between the query and the documents. A 'System' refers to the physical implementation of an IR algorithm, which can have various operational modes or various settings of parameters. Therefore, the same IR system may be used to execute different IR schemes by adjusting the various parameters involved in it.

Performance of the models and strategies differ from corpus to corpus. An early literature[4] indicates that, the existing schemes are non-omnipotent. Hence, it is essential to establish a process that yields reliable and consistent results. Fusion is an active area of research that successfully fulfills the above said need. In data fusion, the data from the multiple sources are merged together and the fusion function 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.

This study focuses on a method that adaptively selects the top-m retrieval schemes. The tasks associated with of the proposed technique of three folded: It has to select best combination function, top-m retrieval schemes and suitable weights for the selected schemes. The above-mentioned functions are carried out using GA.

THE GENETIC ALGORITHM

Genetic Algorithms are inspired by the principle of Darwin's theory of natural selection and survival of the fittest. It is a stochastic search or optimization technique, which is used to find out the optimal solution in a complex state space. It finds an extensive application in Web search, Manufacturing methods and robotics etc. It adapts the natural process such as Reproduction, Crossover and Mutation as it's basic operators.

The Selection process in GA is fitness proportionate one. The individuals present in a generation possess some fitness value. The highly fit, most dominating individual will pass on to the next generation (Survival of the Fittest, Exploitation). The GA uses Crossover and Mutation, as it's primary tools for exploration. These two processes depend on some probability value called as crossover and mutation probability (p_c and p_m). The crossover and mutation operations disturb the string

Corresponding Author: N.P .Gopalan, Department of Computer Applications, NIT, Tiruchirappalli-15, India

structure of the chromosome and producing the new ones. The newborn individuals facilitate the exploration process.

GA maintains the population of points in a solution space and uses an implicit parallelism while analyzing more number of individuals than those present in a single generation. Though, there are various search/optimization tools are available, GA is an active area of research and extending it's application to a wide range of areas like robotics, pipeline construction, etc for the past two decades.

FUSION TECHNIQUES

Fisher[5] employed data fusion for information retrieval by combining together two Boolean searches. His method is confined only to two sources. A linear combination method for fusing multiple sources by assigning weights to the individual strategies was studied by Belkin and Croft[6,7]. The final relevance score assigned using the weighted linear combination method for a document is given by

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

Where,

θ_i: Weight of the i^{th} retrieval strategy,

$E_i(q,d)$: Relevance score returned by the ith retrieval strategy and

k: Number of retrieval strategies to be fused.

The weighted linear combination method has the limitation of requiring prior knowledge about the retrieval systems to assign the weights.

The Comb-functions for combining scores that treat all strategies equally have been proposed by Fox and Shaw[8,9]. The various Comb-functions used for combining scores are shown in Fig. 1. Lee[10,11] has carried out extensive work on Comb-functions. Lee has proposed new rationales and indicators for data fusion. He has conducted experiments over TREC data collection and it was concluded that functioning of CombMNZ is better than the remaining.

The training data for the fusion operation are used to select the best functioning scheme with appropriate weights. Probabilistic approach is used for this purpose. The strategy with best performance is selected automatically from the pool of retrieval strategies based on the predicted probability value in spite of the appreciable performance of the remaining ones. Bilhart[12] proposed a heuristic data fusion algorithm, which uses

CombMIN	Minimum of Individual Similarities
CombMAX	Maximum of Individual Similarities
CombSUM	Summation of Individual Similarities
CombANZ	CombSUM ÷ Number of non zero Similarities
CombMNZ	CombSUM × Number of non zero Similarities

Fig. 1: Comb-functions for combining scores

Genetic Algorithm (GA) for combining the retrieval scores, which assigns weights to independent retrieval strategies and selects the significant ones for fusion.

Applications of Evolutionary algorithm in the area of IR and the usefulness of the GA to i) Indexing, ii) Clustering, iii) Query definition, iv) Matching function learning, v) Image retrieval and vi) User profile updating have been described in detail by cordon[13].

COMBINATION FUNCTIONS AND RETRIEVAL STRATEGIES

In the presented study, four combination functions are used: C-maxmax, C-maxmin, C-minmax and C-minmin[14]. The functions used to assign the final relevance score to the documents are given in the Fig. 2 and are labeled as 0, 1, 2 and 3, respectively.

Retrieval strategies: The similarity measures of Vector Space Model (VSM) and P-norm model with p value 1.5 and 2.5 are chosen as the retrieval strategies. The similarity measure of VSM and the P-norm model is given in the given Eq. 2-4.

$$\text{Inner product } S(q,d) = \sum_{t \in q \cap d} w_{q,t} * w_{d,t} \qquad (2)$$

$$\text{Dice coefficient } S(q,d) = \frac{2 \sum_{t \in q \cap d} w_{q,t} * w_{d,t}}{W_q^2 + W_d^2} \qquad (3)$$

$$\text{P-norm } S(q_{and}, d_j) = 1 - \left(\frac{(1-w_1)^p + (1-w_2)^p + \ldots\ldots + (1-w_m)^p}{m} \right)^{1/p} \qquad (4)$$

Where,

$w_{d,t}$: weight of a term in the document,

$w_{q,t}$: weight of a term in the query,

Wq, Wd-weight of the query and document respectively,

Wm-weight of the m^{th} term in P-norm model and

m-Number of terms present in the corpus for the P-norm model.

$$FRS_{C-max\,max} = \max_{\forall j,j\neq k}\left(\max_{i=1,2..n}\left(s_i^k - s_i^j\right)\right)$$

$$FRS_{C-max\,min} = \max_{\forall j,j\neq k}\left(\min_{i=1,2..n}\left(s_i^k - s_i^j\right)\right)$$

$$FRS_{C-min\,max} = \min_{\forall j,j\neq k}\left(\max_{i=1,2..n}\left(s_i^k - s_i^j\right)\right)$$

$$FRS_{C-min\,min} = \min_{\forall j,j\neq k}\left(\min_{i=1,2..n}\left(s_i^k - s_i^j\right)\right)$$

Where,
j, k = Document identifier,
I = retrieval strategy and
s_i^j = relevance score of j^{th} document in i^{th} retrieval strategy.

Fig. 2: C-functions for combining scores

The above-mentioned retrieval schemes are labeled as 0, 1, 2 and 3, respectively.

Proposed selection scheme: The intention of combining more sources is to merge the merits of all. Since the functioning of the fusion operation depends on the participating strategies, it is necessary to properly select the members. In the proposed work, GA is used as a search mechanism to find the optimal solution. The role of GA in the selection method is of three fold. First it has to find out the best combination function from the pool. Second it has to select best-m schemes and at last the optimal weights for the selected schemes. For these, GA uses the relevant judgment of a query as the training data and by using it, the selection mechanism adaptively choose the optimal solution.

Fitness function: The non-interpolated average precision value is used as the fitness function and it is given by (5)

$$Fit(f) = \frac{\sum_{i=1}^{n} ave\ p_s(q_i)}{n} \qquad (5)$$

Where,
n-number of queries and
ave p_s-average precision of the query.

In order to eliminate the premature convergence, Fitness Scaling mechanism is used. The scaling operation maps the original raw fitness value (f) to the new one (f'). Linear fitness scaling is chosen as a scaling method and it is given by

$$Scaled\ Value\ \ f' = a.f_{raw} + b$$

$$a = \frac{2.f_{ave} - f_{ave}}{f_{max} - f_{ave}}$$

$$b = \frac{fave.(f_{max} - 2.f_{ave})}{f_{max} - f_{ave}}$$

Where,
f_{raw}: raw fitness value.
f_{ave}: average fitness value and,
f_{max}: maximum fitness value.

String coding: The Binary coding method is used code the chromosomes. The individuals present in the population representing the best combination function, top-m retrieval strategies and weights for the selected retrieval strategies. The number of bits used to represent the string is depends on the `m' value. In the proposed method, value of `m' is chosen as 2, 3 and the corresponding number of bits used to encode the string are 11 and 13, respectively. The string structures for the both cases are explained below.

m = 2
no of bits 11: b = <b_0,b_1,.....,b_{10}>
<b_0,b_1,b_2>-Weight of the First Retrieval Strategy (We1)
We1 = $4\times b_0 + 2\times b_1 + b_2$
<b_3,b_4,b_5>-Weight of the Second Retrieval Strategy (We2)
We2 = $4\times b_3 + 2\times b_4 + b_5$
<b_6,b_7,b_8>-Retrieval Schemes to be selected (Sh)
Sh = $4\times b_6 + 2\times b_7 + b_8$
Sh = 0 → Schemes 0 and 1
Sh = 1 → Schemes 0 and 2
Sh = 2 → Schemes 0 and 3
Sh = 3 → Schemes 1 and 2
Sh = 4 → Schemes 1 and 3
Sh = 5 → Schemes 2 and 3
Sh 6 and 7 are unused. If a string with Sh value 6 or 7 is encounter in a generation due to crossover and mutation operation, fitness value of 0 is fixed for that corresponding string.
<b_9,b_{10}>-Combination Function (Cf)
Cf = $2\times b_9 + b_{10}$
Cf = 0 → Combination Function 0
Cf = 1 → Combination Function 1
Cf = 2 → Combination Function 2
Cf = 3 → Combination Function 3

m = 3
no of bits 13: b = <b_0,b_1,.....,b_{13}>

<b_0,b_1,b_2>-Weight of the First Retrieval Strategy (We1)
We1 = $4\times b_0 + 2\times b_1 + b_2$
<b_3,b_4,b_5>-Weight of the Second Retrieval Strategy (We2)
We2 = $4\times b_3 + 2\times b_4 + b_5$
<b_6,b_7,b_8>-Weight of the Third Retrieval Strategy (We3)
We3 = $4\times b_6 + 2\times b_7 + b_8$
<b_9,b_{10}>-Retrieval Schemes to be selected (Sh)

$Sh = 2 \times b_9 + b_{10}$

$Sh = 0 \rightarrow$ Schemes 0, 1 and 2

$Sh = 1 \rightarrow$ Schemes 0, 1 and 3

$Sh = 2 \rightarrow$ Schemes 0, 2 and 3

$Sh = 3 \rightarrow$ Schemes 1, 2 and 3

$<b_{11},b_{12}>$-Combination Function (Cf)

$Cf = \$2 \times b_9 + b_{10}$

$Cf = 0 \rightarrow$ Combination Function 0

$Cf = 1 \rightarrow$ Combination Function 1

$Cf = 2 \rightarrow$ Combination Function 2

$Cf = 3 \rightarrow$ Combination Function 3

RESULTS

The experiments are conducted on the three benchmark test collections namely ADI, CISI and MED. All the queries present in each of the collections used for training purpose.

GA parameters: The number of generation is fixed at 100. The crossover and mutation probability (P_c and P_m) are chosen as 0.6 and 0.01 respectively. In the experiment elitism is used. A highly fit individual from each generation will pass on to the next generation without undergoing crossover and mutation. Roulette Wheel method[15] is used for the selection. In order to over come the draw back of the single point crossover (poised towards the end) two-point crossover is used. The number of individual present in a generation is chosen as 10. The Table 1 summarizes the GA parameters used in the experiment.

Algorithm: The algorithm explains the process of finding the optimal solution in a controlled manner. The steps involved in finding the solution are given below.

{
Initialize the number of generations g
Initialize the number of chromosome n
Randomly generate the initial population p(g)
Evaluate p(g)
While (non termination condition)
do
{
g=g+1
Reproduce n-1 chromosomes p(g) from p(g-1)
Select a highly fit string form p(g) and pass it to the next generation g+1
Recombine p(g)
Mutate p(g)
}
}

Table 1: GA parameters

No. of generations	100
No. of individuals	10
Crossover probability p_c	0.6
Mutation probability p_m	0.01

Table 2: Average precision value

ADI		
Best Single Scheme		0.4243
Combination	Precision	Improvement
Best 2 Combination	0.4572	7.73
Best 3 Combination	0.456	7.47
MED		
Best Single Scheme	0.4544	
Best 2 Combination	0.4854	6.8
Best 3 Combination	0.4963	8.4
CISI		
Best Single Scheme	0.2205	
Best 2 Combination	0.2386	8.2
Best 3 Combination	0.2328	5.57

Table 3: T Value

ADI	
2 Combination	7.756
3 Combination	5.547
CISI	
2 Combination	2.639
3 Combination	3.511
MED	
2 Combination	8.817
3 Combination	7.306

The 'elitism' is used to exploit the highly fit individual and it maintaining a history but the period of the history is restricted to one generation in our case.

The non-interpolated average precision value is used as the performance indicator in the experiment. The Table 2 shows the value of the precision for the best single strategy, best 2 combinations and the 3 combination obtained over the three-benchmark test collections rounded of to 4 decimal place.

The Table gives the maximum value for the precision and the corresponding % of improvement obtained in the experiment over the number of trials with out changing the parameters. The Fig. 3 shows the 11 pt interpolated precision curve to simplify the comparison process.

Test of significance: The experiment is conducted over a number of times and the maximum value obtained is presented in the Table 2. Even though the results look very promising, we planed to test it's significance. For the above said purpose, student-t test is used and the number of trial for the experiment is fixed at 30. The indention of the significance test is to check whether the proposed method yields consistent performance. The hypothesis used for the test is given below.

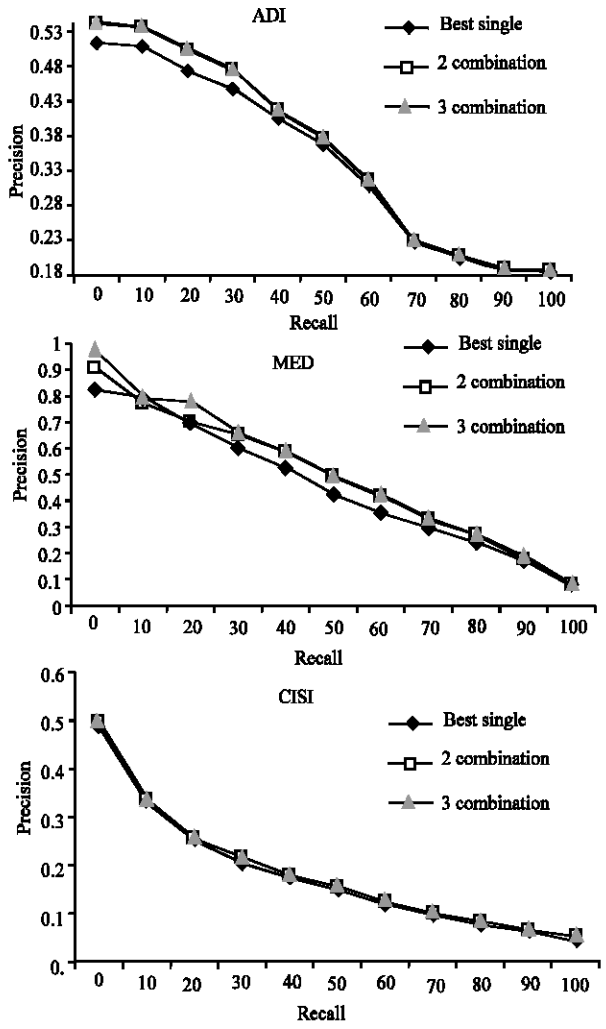

Fig. 3: 11-Point Interpolated precision for best single, two and three combination schemes

H_0: $\mu \leq$ Average non-interpolated precision value of the best single retrieval scheme

H_1: $\mu >$ Average non-interpolated precision value of the best single retrieval scheme

The null hypothesis is successfully rejected as per the calculated t value, which is shown in the Table 3.

The value of `t' for the best 2 and 3 combination over the 3 test collections in the table indicates that, the proposed Adaptive Selection method significantly yields the better and consistent results than the best single strategy participating in the selection process. The proposed selection process is proved to be a better one at 1% standard error level.

CONCLUSION

The proposed adaptive selection method significantly improves the performance of the fusion system. The experiments are conducted over a medium sized test collection. As GA passes a highly fit individual to the next generation, it is indented to test the impact of number of individuals to be passed to next generation as well as the history duration i.e. the number of generations to be considered for selecting the highly fit individual.

REFERENCES

1. Korfhage, R.R., 1997. Information storage and retireval. Willey Computer Publishing.
2. Salton, G. and M.J. McGill, 1983 Introduction to modern information retrieval. McGraw-Gill.
3. Yates, R.B. and B.R. Neto, 1999. Modern information retrieval. Pearson Education.
4. Zobel, J. and A. Moffat, 1988. Exploring the Similarity space. ACM SIGIR Forum, 32: 18-34.
5. Fisher, H.L. and D.R. Elchesen, 1972. Effectiveness of combining title words and index terms in machine retrieval searches. Nature, 238: 109-110
6. Croft, B., 2000. Combining Approaches to Information Retrieval. In Advances in Information Retrieval. Edited by W.B.Croft. Kluwer Academic Publishers, pp: 1-36
7. Cool, C., N. Belkin, P. Kantor and R. Quatrain, 1994. Combining Evidence for Information Retrieval. In Proceedings of the 2nd Text Retrieval Conference, pp: 35-44.
8. Fox, E.A. and J.A. Shaw, 1994. Combination of Multiple Searches. In proceedings of the Second Text Retrieval Conference, pp: 243-252
9. Fox, E.A. and J.A. Shaw, 1995. Combination of Multiple Searches. In proceedings of the Third Text Retrieval Conference, pp: 105-108
10. J.H. Lee, 1995. Combining Multiple Evidence from Different Properties of Weighting Schemes. In Proceedings of the 18th Annual International Acm Sigir Conference on Research and Development in information retrieval, pp: 180-188
11. Lee, J.H., 1997. Combining Multiple Evidence from Different Relevant Feedback Networks. In Proceedings of the 5th International Conference on Database Systems for Advanced Applications, pp: 421-430.

12. Bilhart, H., 2003. Learning retrieval expert combinations with genetic algorithm. Intl. J. Uncertainty, Fuzziness and Knowledge-Based Sys., 11: 87-114.
13. Herrea-Viedma, E. and C. Cordon, 2003. A review on the application of evolutionary computation to information retrieval. Intl. J. Approximate Reason., 34: 241-264.
14. Gopalan, N.P. and K. Batri, 2006. An effective pareto optimality based fusion technique for information retrieval. Accepted for publication in Vivek: An Artificial Intelligence Journal.
15. David, E. Goldberg, 1989. GA in search, optimization and machine learning. Addison Wesley publication.