Effective Feature Classification of Information Retrieval Schemes for Clustering

[1]Dr. S. Meenakshi Sundaram, and [2]U. Revathy
[1]Associate professor- Department of CSE, Pannai College of Engineering and Technology, Sivagangai, India
[2]Assistant professor- Department of Computer Science, Government Arts College for women, Sivagangai, India

Abstract: Effective information retrieval is defined as the number of relevant documents that are retrieved with respect to user query. Feature selection plays an important role in combining the results from two or more search engines into a unified top-ranked relevant documents list in the context of user information need. The classification algorithm include support vector machine (SVMs) with several kernels and k-nearest neighbor (k-nn). In this paper, we present a novel data fusion in IR to enhance the performance of the retrieval system. The study shows that our approach is more efficient and suitable for IR researchers.

Introduction

A retrieval system is a machine that receive the user query and generate the relevance score for the query-document pair. The process of finding the needy information from a repository is a non-trivial task [1,2,3] and it is necessary to formulate a process that effectively submits the pertinent documents. The process of retrieving germane articles [4] is termed as Information Retrieval (IR). It deals with the representation, storage, organization of and access to the information items [3]. Fusion is a technique that merge results retrieved by different systems to form a unique list of documents. Document Clustering is based on particular ranked list and does not take benefit of multiple ranked list. The objective of clustering is to split the relevant documents from non-relevant documents. The fusion function accepts these score as its output for the query document pair. A static fusion function has only the relevance scores for a single query-document pair as its inputs. A dynamic fusion function can have more inputs. To construct a dynamic fusion function that can adjust the way it fuses multiple retrieval systems relevance scores for a query document pair using additional input features such as query, retrieved documents and joint distribution of retrieval systems relevance score for the query. Various models, schemes and systems have been proposed to represent and organize the document collection in order to reduce the users' effort towards finding relevant information [5].In this paper , different combination of feature selection techniques and classification algorithms for feature classification were studied. The classification algorithm including SVMs with several kernels and K-nn. Finally, an unknown feature can use the trained classification algorithm for the prediction and the classification algorithm will predict the features as relevant or irrelevant.

Related Work

Fox and Shaw showed the five combination function for combining scores[6]. They are as follows:

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.

Fusion functions which are different from Comb-functions with respect to the generation of answer sets, are also found in the literatures [8]. These functions assign ranks to the documents in the answer set against the relevance score assignment mechanism adapted in Comb-functions. Few such fusion techniques which emulate the social voting schemes, are the Borda and Condorcet fusions [8]. Extensive work on Comb functions has been carried out by Lee [9–11] and based on the results he proposed few new rationales and indicators for data fusion. He concluded that CombMNZ is the better performing function than the others.

Guttmans Point Alienation

The pair wise similarity measure include: The number of documents in the intersection of the two lists of returned documents (I)

The correlation coefficient from a linear regression of the scores of documents in the intersection of the two systems (C), which is actually the r^2 value of a regression which uses one system's scores to predict the other's .

The number of relevant documents returned by one system but not the other divided by the total number of relevant documents returned by that system

$$GPA = \frac{\sum_{x,x'} \left(\rho_1(x,q) - \rho_1(x',q)\right) \left(\rho_2(x,q) - \rho_2(x',q)\right)}{\sum_{x,x'} |\ \rho_1(x,q) - \rho_1(x',q)||\rho_2(x,q) - \rho_2(x',q)|}$$

Lees Overlap Measure

Lee's [Lee, 1997] overlap measures, O_{rel} and O_{nonrel}, which measure the proportion of relevant and nonrelevant documents in the intersection of the two lists. These two measures are calculated as:

$$O_{rel} = \frac{2 * I_{rel}}{R_1 + R_2}$$

$$O_{nonrel} = \frac{2 * I_{nonrel}}{N_1 + N_2}$$

where R_i is the number of relevant documents and N_i is the number of nonrelevant documents returned by the system i respectively. The ratio of the two systems found to be an important predictive factor for the improvement of the combination. The similarity measure is the two systems on relevant document is less important than on relevant ones. After normalizing the scores for each system on each query by dividing their respective means we found the optimal combination for each possible. For each feature, we use one of the statistical methods such as the traditional t-test. Large score suggests that the corresponding feature has different expression levels in the relevant and irrelevant documents and thus is an important feature and will be selected for further analysis. Besides that some researchers used a variation of correlation coefficient to select features, for example Fisher Criterion [13] and Golub Signal-to-Noise.

Clustering Hypothesis

our method is based on clustering hypothesis:

Clustering Hypothesis: A modest clustering algorithm is used to split relevant documents from non-relevant documents.

Re-ranking

After clustering each ranked list , the resultant group of clusters each of which contains relevant and non-relevant documents. By re-ranking , we expect to determine reliable clusters and adjust the relevance score of documents in each ranked list such that the relevance scores become more reasonable. To identify reliable clusters, we assign each cluster a reliability score. According to the Fusion hypothesis, we use the overlap between clusters to compute the reliability of a cluster. The reliability of cluster is computed as follows

Symbol	***Explanation***
Q	A query
d	A document
RL_A,RL_B	Ranked list returned by retrieval system A and B, respectively
$C_{A,i}$	i th cluster in RL_A
$sim\_CC(C_{A,i}, C_{B,j})$	Similarity between
$sim\_qC(q, C_{A,i})$	Similarity between query q and
$r(C_{A,i})$	Reliability of cluster
$rel_A(d)$	Relevance score of d given by A
$rel_A^*(d)$	Adjusted relevance score of d
rel(d)	Final relevance score of d

[1] $\sum_j \Big[\frac{sim\_qC(q,C_{B,j})}{\sum_t sim\_qC(q,C_{B,t})} sim\_CC(C_{A,i}$

[1]

[2] $$sim\_CC\left(C_{A,i}, C_{B,j}\right) = |C_{A,}$$ [2]

[3] $$sim\_qC\left(q, C_{A,i}\right) = \frac{\Sigma_{d\epsilon}}{}$$ [3]

In equation 2 , the similarity of two clusters is estimated in terms of similar documents between them. In equation 3 , similarity between query and cluster is estimated in terms of the average relevance score of the documents. In equation 1, for each cluster and all clusters in the ranked list returned by retrieval system B. Two clusters [a and b] from different ranked lists that have the largest overlap are identified to be reliable clusters. Three step approach is used by first clustering each ranked list. After clustering , each ranked list is composed of a set of clusters, say C_1,C_2,.., C_n. Then adjust the relevance value of each document according to the reliability of the cluster. We finally use CombSUM to combine the adjusted ranked lists and present to user. Each reliability represents the precision of a cluster, the below formula adjusts the relevance score of a document in a high reliable cluster.

$$rel_A^*(d) = rel_A(d) * \left[1 + r\left(C_{A,t}\right)\right]$$

Where

Feature Classification

Support Vector Machines

Support Vector Machines are comparatively new category of classification algorithms. An SVM expects a training data set with positive and negative as an input. It then creates a decision limit (the maximal-margin separating limit) between two category and selects the most relevant documents involved in the decision making process (the so called support vectors). The construction of the linear limit is always possible as longs as the document is linearly separable. SVMs can use kernels, which provide a nonlinear mapping to a higher dimensional feature space. The dot product has the following formula:

$$K(x,y) = (x.y + 1)^d$$

Where x and y are the vectors of the text, the parameter d is an integer which decides the partition. In the case where d is equals to 1, a linear classification algorithm is generated and it is called the SVM dot product. In the case where d is more than 1, a nonlinear classification algorithm is generated and it is called SVM quadratic dot product.. In this paper, where d is equals to 3, it is called the SVM cubic dot product. The radial basis kernel is as follows,

$$K(x,y) = exp\left[\frac{|x-y|^2}{2\sigma^2}\right]$$

Where is the median of the Euclidean distances between the members and non members of the document category.

The main advantages of SVMs are that they are robust to outliers, join quickly, and find the optimal decision limit if the document is predictable. Another advantage is that the input document can be mapped into an arbitrary high dimensional document clusters where the linear decision limit can be predicted. This mapping allows for higher order of interactions between the samples and can also find correlations between documents. SVMs are also very flexible as they allow for a big variety of kernel functions.

K-nearest neighbor

The k-nn classification algorithm is a simple algorithm based on a distance metrics between the testing documents and the training documents. The main design of the DCP system is, given a testing document s, and a set of training tuples T containing pairs in the form of where is the values of document I and is the class label of document i. Find k training sample with the most similar documents between t and s, according to distance measure. The class label with the highest rank among the k training sample is assigned to s. The main advantage of k-nn is it has the ability to model very complex target functions by a collection of less complex approximations. It is easy to program and understand. It is robust to noisy training samples.

Discussion

The strength of this approach should be the possibility of picking up optimal solution from different retrieval systems.

Despite all the problems, several information retrieval researchers think that it is necessary to adhere to the standards coming from the TREC, trying to solve possible performance problems. Ranking a group of retrieval systems consists of determining a perfect ordering according to qualified importance of retrieval systems. The work up to now has shown the feasibility of the design. It is in the hands of the researchers to utilize this article and transform it into a reality thereby enhancing the quality parameters, precision and recall values.

Acknowledgment

Our Sincere thanks to the almighty who is behind us in all aspects of developing this paper from various research articles that is transformed into a review paper.

References

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

[2] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw–Gill, 1983.

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

[4] G. W. Cottrell B. T. Bartel and R. K. Belew. “Learning to retrieve information”, In Current Trends in Connectionism: Proceedings of the Swedish Conference on Connectionism, pp. 345–354, 1995.

[5] Swedish Conference on Connectionism, pp. 345–354, 1995.

[6] C. J. Van Rijsbergen. Information Retrieval. Butterworth-Heinemann, 1979.

[7] E. A. Fox and J. A. Shaw.” Combination of multiple searches”. In Proceedings of the Second Text Retrieval Conference TREC 2, pp. 243–252,1994.

[8] M. Montague and J. A. Aslam,” Condorcet fusion for improved retrieval”,.In Proceedings of the `

[9] G. Mauris L. Valet and P. Bolon ”A statistical overview of recent literature in information fusion”, In Procedings of the Third International Conference on Information Fusion, pp. 22–29, July 2000.

[10] Joon Ho Lee,” 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, 1995.

[11] Joon Ho Lee,” Analyses of multiple evidence combination”, In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 260–276, 1997.

[12] Joon Ho Lee,” Combining multiple evidence from different relevant feedback networks”, In Proceedings of the 5th international Conference on Database Systems for Advanced Applications, pp. 421–430, 1997.

[13] Savoy. ‚Combining probabilistic and vector schemes,”In Proceedings of the Fourth Text Retrieval Conference, pp. 537–548, 1996.

[14] C. A. Coelo Coelo. “An updated survey of ga-based multiobjective optimization techniques”, ACM Computing Survey, pp. 109–143, Jan 2000.

[15] H. Bilhart. “Learning retrieval expert combinations with genetic algorithm”, International Journal of Uncertainity,Fuzziness and Knowledge Based Systems, 11 (1):87–114, Feb 2003.Kevin R. Fall, W. Richard Stevens, TCP/IP Illustrated, Volume 1: The Protocols, 2nd ed., Addison-Wesley, USA, 2011.

[16] Jian Zhang, Jianfeng Gao, Ming Zhou, Jiaxing Wang,`` Improving the Effectiveness of Information Retrieval with clustering and Fusion”, To appear in the Computational Linguistics and Chinese Language Processing,

[17] Rabia Nuray and Fazli Can. Automatic ranking of information retrieval systems using data fusion. Information Processing and Management, 42(3):595–614, May 2006.