Ranking of Web Documents Using Multimodal Cross Reference Re-Ranking

Dr. S. Vijayarani[1], E. Suganya[2]
[1]*Assistant Professor*
[2]*Ph.D Research Scholar*
Department of Computer Science
Bharathiar University, Coimbatore- 641046
Tamilnadu, India

Abstract - *The dynamic web has increased exponentially over the past few years with more than thousands of documents related to a subject available to the user now. The growth of web becomes very difficult to get the proper information related to user query. Most of the web documents are unstructured /semi-structured and not in an organized manner and hence user facing more difficult to find relevant documents. When user searches some information on the web which returns huge amount of web pages in response to user queries. It is not possible to one and all to explore all web pages. A more useful and efficient mechanism is ranking the web documents. Web mining helps in retrieving potentially useful information and patterns from web. This paper is we have proposed Multimodal Cross Reference Re-ranking method for ranking the web documents which are extracted from the web. It improves the performance of search engine. This approach will helps the user to get all their relevant documents in one place and can restrict their search to some top documents of their choice.*

Keywords: Web Mining, Ranking, Cross Reference Ranking and MCR Ranking

1. INTRODUCTION

The World Wide Web increasing at an exponential rate rapidly, it is becoming increasingly difficult to find relevant information because it has large amount of information. However, considering the size of the World Wide Web, a typical query might give more than a million results for user query. The user does not have the time or patience to go through this huge list [3]. Thus, ranking of web documents becomes a critical component of information retrieval. The main aim of this paper is to find better and more efficient ranking methods, which can return high quality information to the user in as small a time frame as possible. Search engines first create an index of all the web documents and store it on the server [12]. After the user submits a query, the query is given to the index, which returns the documents containing the words in the query. Then, the returned documents are sent to a ranking function which gives a rank to each document and the top-*k* documents are returned to the user [5].

Web Mining is the task of extracting useful information from web documents. Web Mining comprises of three types: Web Structure Mining, Web Content Mining, and Web Usage Mining. Ranking of query results is one of the fundamental problems in information retrieval. To overcome the problem we have proposed a new technique for ranking the web documents.

The remaining section of the paper is organized as follows. Section 2 discusses the methodology of the proposed system. The experimental results are given in section 3. Section 4 provides the conclusion of this paper.

2. METHODOLOGY

Figure 2.1 shows the overall system architecture of the proposed work.

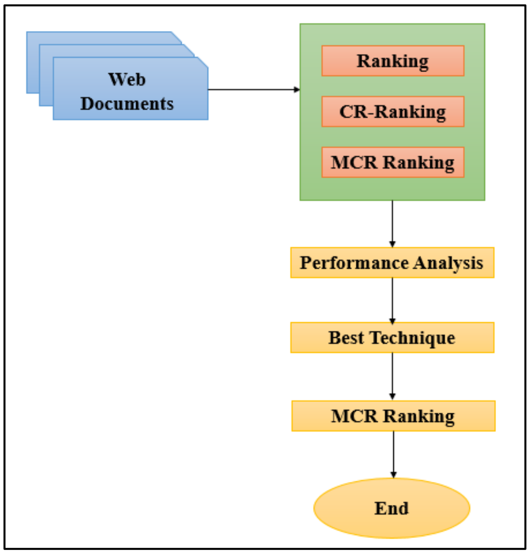

Figure 2.1. System Architecture

2.1 Web Documents

A web page is a document that is suitable for the World Wide Web and web browsers. It has three types namely static, dynamic and active. A static web document resides in a file that it is associated with a web server [3] [4]. The content of the static web documents do not change. A dynamic web document does not exist in a predefined form. When a request arrives the web server runs an application program that creates the document [6]. The server returns the output of the program as a response to the browser that requested the document. Because a new document is created for each request, the contents of a dynamic document can vary from one request to another [8]. An active web document consists of a computer program that the server sends to the browser and that the browser must run locally. When it runs, the active document program can interact with the user and change the display continuously [7].

2.2 Ranking

Ranking is one of the important tasks of information extraction in web mining. Searching the web involves two main stages [15]. Extracting the pages relevant to a query and ranking them according to their quality. Ranking is an essential as it helps the user looks for "quality" pages that are related to the query. Different metrics have been proposed to rank web pages according to their quality. With the rapid growth of the Web, providing relevant pages of the highest quality to the users based on their queries becomes increasingly difficult [7]. The reasons are that some web pages are not self-descriptive and that some links exist purely for navigational purposes. Therefore, finding suitable pages through a search engine that relies on web contents or makes use of hyperlink information is very difficult.

Ranking functions are evaluated by a variety of means; one of the simplest is determining the precision of the first *k* top-ranked results for some fixed *k*; for example, the proportion of the top 10 results that are relevant, on average over many queries [14]. Frequently, computation of ranking functions can be simplified by taking advantage of the observation that only the relative order of scores matters, not their absolute value; hence terms or factors that are independent of the document may be removed, and terms or factors that are independent of the query may be pre-computed and stored with the document.

2.3 Cross Reference Ranking

Re-Ranking is a method to merge features in the method of cross reference, which combines the features in the manner of cross reference [16] [17]. This method is to utilize for inferring the most relevant web documents in the initial search results. CR Re-ranking method contains three main stages: Specifically, the initial search results are first divided into several clusters individually in different feature spaces. Then, the clusters from each space are mapped to the predefined ranks according to their relevance to the query [18]. Given the ranked clusters from all the feature spaces, the cross-reference strategy can hierarchically merge them into a unique and improved result ranking.

2.4 Multimodal Cross Reference Re-Ranking

In this work we have proposed Multimodal Cross Reference Re-ranking method for ranking the documents which are extracted from the web [1]. It improves the performance of search engine. The results are processed individually in two different feature views, i.e., feature A and feature B. In each feature view clustering is performed. To attain three clusters in feature view A and feature view B. Then these clusters are ranked as High, Medium and Low, according to their relevance to the query. Finally, a unique and improved result set is formed by hierarchically combining all the ranked clusters from two different views [2]. After extracting multiple features for each document, carry out clustering independently in these features views. This provides a possibility for offering high accuracy on top ranked documents [9]. As a result, to obtain a certain number of clusters from each feature view, which gives the way for implementing cross reference strategy.

MCR-Re-ranking Algorithm

1. Initial Result is taken and it is processed in two distinct feature views, i.e. feature A and feature B

2. In each feature view they are ranked in ascending order based on Euclidean Distance

Let A= {A_1, A_2... A_{10}}

$$\boxed{md(A_i, A\backslash A_i) = \min_{A_j \in A\backslash A_i} \{d(A_i, A_j)\},}$$

d(.,.) is the Euclidean distance, **md** is the smallest distance possible

3. In each feature view all the results are first clustered into three clusters and then they are mapped into three predefined rank levels i.e., High, Medium, and Low based on their relevance to the query.

4. All the ranked clusters, from the different features are hierarchically combined using cross reference strategy.

5. Two ranked clusters can be integrated into a unique ranked subset list using the rule:

Rank ($\mathbf{A}_{\text{high}} \cap \mathbf{B}_{\text{medium}}$) > Rank ($\mathbf{A}_{\text{medium}} \cap \mathbf{B}_{\text{low}}$)

If (high + medium) < (medium + low)

$\mathbf{A_{high}}$, $\mathbf{A_{medium}}$ are the clusters of feature view A and $\mathbf{B_{medium}}$, $\mathbf{B_{low}}$ are the clusters of feature view B.

6. When (high + medium) = (medium + high), using Hausdorff distance as follows:

   Rank ($\mathbf{A_{high}} \cap \mathbf{B_{medium}}$) > Rank ($\mathbf{A_{medium}} \cap \mathbf{B_{high}}$)

   When **hd** (E, $\mathbf{A_{high}} \cap \mathbf{B_{medium}}$) < **hd** (E, $\mathbf{A_{medium}} \cap \mathbf{B_{high}}$)

   where E is the query relevant set.

7. Thus a final result set is formed and accuracy is achieved on top ranked results.

After getting different clusters from each feature view, the next step is to rank them in accordance with their relevance to the query given by the user. Some query relevant documents should be selected in advance to convey the intent of the query. Hence top ranked initial results are considered as informative documents. Ranking is done in ascending order according to the following distance:

$$\boxed{md(a_i, A\backslash a_i) = \min_{a_j \in A\backslash a_i}\{d(a_i, a_j)\},}$$ [11]

Here d(.,.) is the Euclidean distance and ai and aj are calculated. The distance between relevant documents is smaller when compared to those distances between irrelevant documents or between relevant documents and irrelevant documents [18]. Hence, relevant documents are grouped together and irrelevant documents are scattered. Consider E as query relevant document. 'K' documents with the smallest distances can be the most possible documents that convey the intent of the query. To measure the relevance between documents sets, to employ the modified Hausdorff distance [9], which is defined as follows:

$$\boxed{hd(E, C) = mean_{e \in E}\{\min_{c \in C}\{d(e, c)\}\},}$$

Here E is the query-relevant set and C can be a cluster or any document set. hd (E,C) is a directed Hausdorff distance from E to C. Assign corresponding ranks to the clusters in each feature view. Here, considered three ranks for the clusters in each feature view. The three ranks are High, Medium and Low.

The final goal of this proposed method is to get high accuracy on the top ranked documents, by using an improved re-ranking on the initial results. Thereby in order to move in the direction of this goal, hierarchically combine or merge all the clusters that are ranked in different feature views in the previous step. The cross reference method is used to merge all the clusters of one feature view with the clusters of another feature view. There are many clustering algorithms for document clustering. Our task is to cluster a small collection of documents returned by individual retrieval systems.

Ranking within the Cluster

1. Randomly set document **di** to cluster **Cj**
2. LoopCount =0; ShiftCount = 1240;
3. **While** (LoopCount < LoopThreshold and ShiftCount > ShiftThreshold)
4. **Do**
5. Construct the centroid of each cluster, **i.e.**

   $$\text{Centroid of } C_j = \frac{\sum_{d_i \in C_j} d_i}{|C_j|}$$

6. Assign **di** to its nearest cluster(the distance is determined by the similarity between **di** and the centroid of cluster);
7. ShiftCount = the number of documents shift to other cluster;
8. LoopCount ++;

Experimental results show that the search effectiveness, especially on the top-ranked results, is improved significantly.

3. EXPERIMENTAL RESULTS

In this method, web documents are considered as the fundamental unit. Hence, feature extraction is based on web documents. For each webpage, two features are extracted: Link and HTML schemas. The performance evaluation includes precision at different depths of result list (Precision D), non-interpolated average precision (AP), and mean average precision (MAP). D denotes the depth where precision is computed. Let S be the total number of returned documents and R_i the number of true relevant documents in the top-i returned results. The evaluation criteria can be defined as follows:

$$\boxed{Prec_{D(T_n)} = \frac{1}{D}\sum_{i=1}^{D} F_i}$$ [18]

$$\boxed{AP(T_n) = \frac{1}{R}\sum_{i=1}^{S}(\frac{R_i}{i}.F_i)}$$ [18]

$$\boxed{MAP = \frac{1}{N}\sum_{n=1}^{N} AP(T_n)}$$ [18]

Where T_n is the *n*th query topic, F_i=1 if the *i*th document is relevant to the query and 0 otherwise, R stands for the total number of true relevant documents, N denotes the number of query topics. Precision D is utilized to assess the precision at the different depth of the result list. AP shows the performance of a single query topic, which is sensitive to the entire ranking of the documents. MAP summarizes the overall performance of a search system over all the query topics. Table 3.1 and Figure illustrate the performance evaluation of proposed re-ranking method.

Table3.1 Performance of Proposed Re-ranking Method

Method	Precision (10 docs)	Precision (20 docs)	Precision (30 docs)	AP	MAP (%)
Ranking	0.132	0.112	0.013	0.086	10
CR Ranking	0.144	0.124	0.157	0.141	36
MCR Ranking	0.201	0.172	0.201	0.191	41

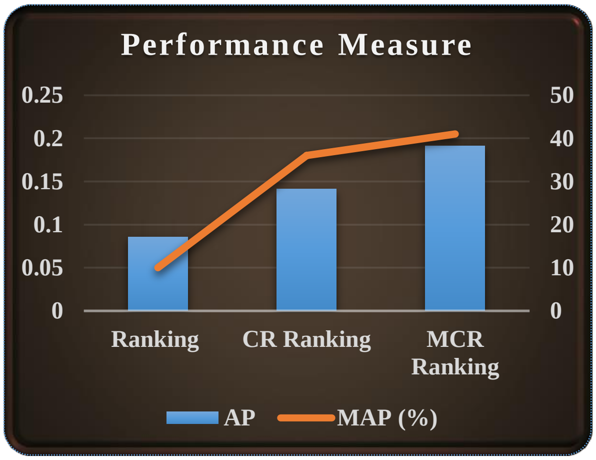

Figure 3.2 Performance of Proposed Re-Ranking Method

4. CONCLUSION

In this paper, we have proposed a new method for ranking the web documents, which are extracted from the web. It improves the performance of search engine. This approach will helps the user to get all their relevant documents in one place and can restrict their search to some top documents of their choice. The proposed ranking method deals with improving the search strategies several ways, thus retrieves the most relevant pages. This ranking method takes the concepts and relationship between the concepts which exists both in the document and user query to improve the retrieval of relevant document.

REFERENCES

[1] Nisha, Dr. Paramjeet singh, "A Review Paper on SEO based Ranking of Web Documents", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 7, July 2014.

[2] Rajendra Kumar Roul, Omanwar Rohit Devanand, S. K. Sahay, "Web Document Clustering and Ranking using Tf-Idf based Apriori Approach"

[3] Poonam Chahal, Manjeet Singh, Suresh Kumar, "Ranking of Web Documents using Semantic Similarity", International Conference on Information Systems and Computer Networks.

[4] Shashank Gugnani, Tushar Bihany, Rajendra Kumar Roul, "A Complete Survey on Web Document Ranking", International Journal of Computer Applications, Volume ICACEA - No. 2, 2014.

[5] Donna Harman, "Ranking Algorithms", Information Retrieval: Chapter 14.

[6] Michael Bendersky, W. Bruce Croft, Yanlei Diao, "Quality-Biased Ranking of Web Documents"

[7] Zakaria Suliman Zubi, "Ranking Web Pages Using Web Structure Mining Concepts", Recent Advances in Telecommunications, Signals and Systems.

[8] Rekha Jain, Dr. G. N. Purohit, "Page Ranking Algorithms for Web Mining", International Journal of Computer Applications, Volume 13– No.5, January 2011.

[9] A.M. Sote, Dr. S. R. Pande, "Application of Page Ranking Algorithm in Web Mining", IOSR Journal of Computer Science.

[10] Vidya Kannan, Dr. G.N Srinivasan, "Yet another way of Ranking web Documents Based on Semantic Similarity", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 4, April 2014.

[11] V. Anthoni sahaya balan, S. Singaravelan, D.Murugan, "Combined Cluster Based Ranking for Web Document Using Semantic Similarity', IOSR Journal of Computer Engineering, Volume 16, Issue 1, Ver. IV , Jan. 2014.

[12] Jiyin He, Krisztian Balog, Katja Hofmann, Edgar Meij, Maarten de Rijke, Manos Tsagkias, Wouter Weerkamp, "Heuristic Ranking and Diversification of Web Documents".

[13] https://www.semanticscholar.org/paper/An-effective-web-document-clustering-for-Roul-Sahay/157bd379a8aecb457f3f01db3028b9a5b6678428

[14] https://arxiv.org/abs/1406.5617

[15] https://en.m.wikipedia.org/wiki/Ranking_(information_retrieval)

[16] P. Perumal, D. Anandhu, "Video Search Reranking Via Cross Reference Based Fusion Strategy", International Journal of Trend in Research and Development, Volume 2(4).

[17] Ravi Regulagadda , G.Yedukondalu, "Video Search Reranking using Multimodel fusion Technique", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 559– 562 (2014).

[18] Priyanka B. Kamdi, Pravin Kulurkar, "Data Mining Approach for Image Retrieval in Multimodal Fusion Using Frequent Pattern Tree", International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 7, July 2015.

BIOGRAPHIES

Dr. S. Vijayarani has completed MCA, M.Phil and Ph.D in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. E. Suganya has completed M.Phil in Computer Science. She is currently pursuing her Ph.D in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining and Web Mining.