ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

Effective Information Retrieval Using Clustering Technique

SK. Saida, CH. Ravindra Reddy

Assistant Professor, Department of Computer Science & Engineering, Sree Vahini Institute of Science & Technology , Tiruvuru, A.P, India
Assistant Professor, Department of Computer Science & Engineering,, Sree Vahini Institute of Science & Technology , Tiruvuru, A.P, India

ABSTRACT: Locating interesting information is one of the most important tasks in Information Retrieval (IR). The different IR systems emphasize different query features when determining relevance and therefore retrieved from different sets of documents. Clustering is an approach to improve the effectiveness of IR. In clustering, documents are clustered either before or after retrieval. The motivation of this paper is to explain the need of clustering in retrieving efficient information that closely associates documents which are relevant to the same query. Here IR framework has been defined which consists of four steps (1) IR system similarity measure (3) document clustering and (4) ranking of clusters. Furthermore, we present the short comings of cluster algorithms based on the various facets of their features and functionality. Finally based on the review of the different approaches we conclude that although clustering has been a topic for scientific community for three decades, there are still many open issues that call for more research.

KEY WORDS: Information Retrieval, IR System, Document Clustering, Similarity measure, Ranking

I.INTRODUCTION

The purpose of information Retrieval is to store documents electronically and assist user to effectively navigate, trace and organize the available web documents [16]. The IR system accepts a query from the user and responds with a set of documents. The system returns both relevant and non-relevant material and a document organization approach are applied to assist the user in finding the relevant information in the retrieved set. Generally, a search engine presents the retrieved document set as a rank list of documents. The documents in the list are ordered by the probability of being relevant to the user's request. The highest ranked document is considered to be the most likely relevant document; the next one is slightly less likely and so on. Every search engine works on above organizational approach. The user will start at the top of the list and follow it down examining the documents one at a time. A number of alternative document retrieval approaches have been developed over the recent years [1,8,11,16]. These approaches are normally based on visualization and presentation of some relationships among the documents and the user's query.

The goal of clustering is toseparate relevant documents from non-relevant documents. To accomplish this we define a measure for similarity between documents and design corresponding clustering algorithm. We can start with vector space model(VSM), which represents a document as a vector of the terms that appear in all the document set. Each feature vector contains term weights of the terms appearing in that document. The term weighting scheme is usually based on tf×idf method in IR. A collection of documents can be represented by a term-document matrix. A similarity between documents is measured using one of several similarity measures that are based on relations of feature vectors. After clustering algorithms the clusters are ranked using the ranking algorithm, which generates clusters according to their match with the query. Section 2, of this paper summarizes related work in this area. In Section 3, IR using clustering framework is defined. Section 4, shows the conclusion and discussion

II. SIGNIFICANCE OF THE SYSTEM

Document Clustering is a technique employed for the purpose of analyzing statistical data sets. Essentially, the goal of clustering is to identify distinct groups within a dataset, and then place the data within those groups, according to their relationships with each other [6,22]. Clustering of multidimensional data is an important procedure in many information retrieval applications. In these applications, one or more clustering algorithms are used to group similar items together to form clusters. There exists a large number of data clustering algorithms which are classified into two main categories-Hierarchical algorithms and Partitional algorithms.

III. LITERATURE SURVEY

IR is the act of sorting, searching and retrieving information that matches the user's request. Until 1950's the IR was mostly a library science [16]. Recently, clustering has been used as an alternate organization of retrieved documents, aiming to help users better understand the retrieved documents and therefore be better able to focus their search. The document clustering has been traditionally investigated mainly as a means of improving the performance of IR by pre- clustering the entire corpus [2]. However, clustering has also been investigated as post-retrieval document browsing technique. Numerous document clustering algorithms appear in the literature including K-means [5], hierarchical agglomerative clustering [10], scatter/gather[7]and suffix tree clustering (STC). When using only textual information for clustering [14] has shown that STC outperforms other algorithms but suffix tree based method suffers from largememory requirements and poor locality characteristics. SHOC [5] uses suffix array for phrase extraction and organizes the snippets in a hierarchy via an SVD (Singular Value Decomposition) approach. Lingo [21] uses SVD on a term- document matrix to find meaningful long labels, generates flat clustering result. Zeng [20] re-formalizes the clustering problem as a salient phrase ranking problem. It uses phrases rather than words and that it allows clusters to overlap. A co- occurrence based hierarchical clustering method is used to group search results into hierarchical and overlapping clusters. CoHC outperforms all other clustering algorithms [26]. The most well known clustering methodology can be divided into two methods according to the structure of the group created as a result of clustering: the hierarchical clustering and non- hierarchical clustering method. There are diverse algorithms associated with each methodology

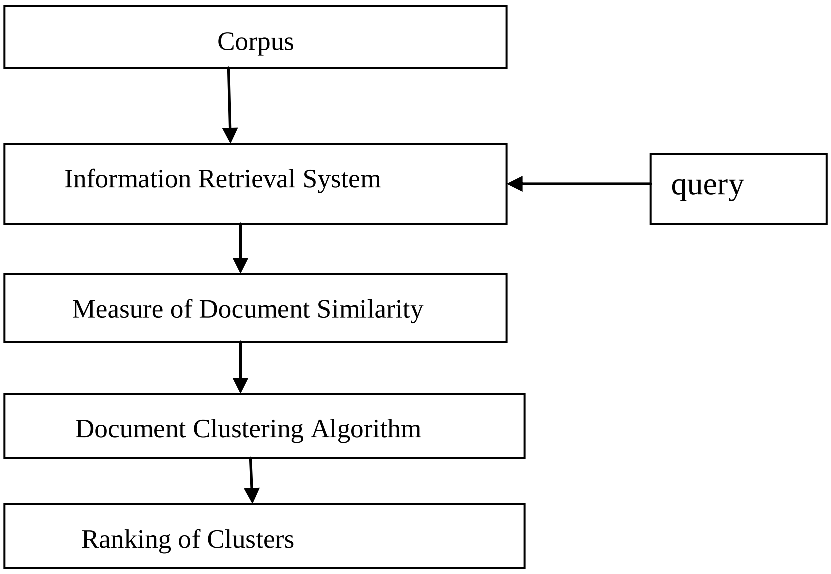
Corpus
Information Retrieval System
query
Measure of Document Similarity
Document Clustering Algorithm
Ranking of Clusters

Figure.1 Information Retrieval Framework overview

A. Information Retrieval System

Today the amount of information available on the Web has increased to a point that there are great demands for effective systems that allow an easy and flexible access to information relevant to specific user's needs [3] The system should be capable of managing imperfect information, and to adapt its behaviour to the user context. Information Retrieval aims at defining models and techniques that improves the limitations of current systems for the Information Access (mainly Information Retrieval and Information Filtering systems).

B. Similarity Measure

Clustering exploits similarities between the documents to be clustered. The similarity of two documents is computed as a function of the distance between the corresponding term vectors for these documents. Of the various measures used to compute this distance, the cosine measure has proved the most reliable and accurate [19].

In order to cluster documents, one must first choose the type and characteristics or attributes of the documents on which the clustering algorithms will be based. The most commonly used model is the Vector Space Model (VSM). The goal of clustering is to separate the relevant documents from the non-relevant documents. To accomplish this we need to define a measure for similarity between documents and appropriate similarity measure must be chosen for the calculation of the similarity between the documents. Some widely used similarity measures are the cosine coefficient which gives the cosine of the angle between the two featured vectors, the Jaccard coefficients, Euclidean and Pearson Correlation and the dice coefficients (all normalised).

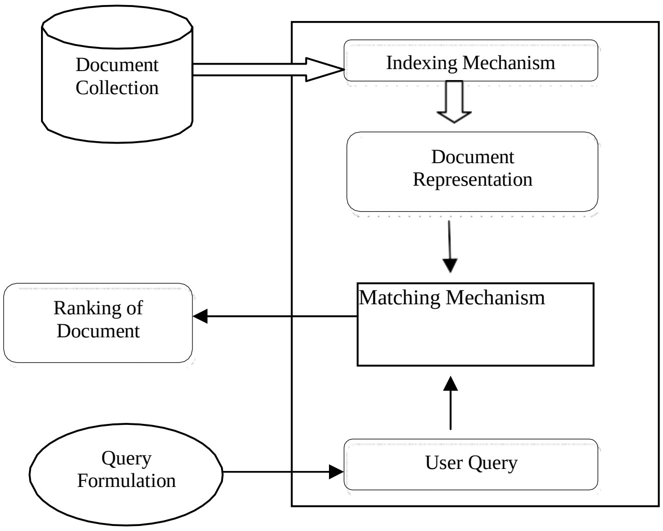

Fig.ure 2. Information Retrieval System

In order to cluster documents, one must first choose the type and characteristics or attributes of the documents on which the clustering algorithms will be based. The most commonly used model is the Vector Space Model (VSM). The goal of clustering is to separate the relevant documents from the non-relevant documents. To accomplish this we need to define a measure for similarity between documents and appropriate similarity measure must be chosen for the calculation of the similarity between the documents. Some widely used similarity measures are the cosine coefficient which gives the cosine of the angle between the two featured vectors, the Jaccard coefficients, Euclidean and Pearson Correlation and the dice coefficients (all normalised).

IV. METHODOLOGY

Clustering-then-labelling approach: Generally, the clustering-then-labelling approach first applies traditional clustering algorithms to group snippets into topically- coherent clusters according to content similarity, and then generates a label for each cluster. However, the cluster labels are often unreadable, which makes it for users difficult to identify relevant clusters [20]. Scatter/Gather system [7,9] is implemented based on a variant of the classic K-Means algorithm. The Scatter/Gather browsing paradigm clusters documents into topically – coherent groups and present descriptive textual summaries to the user. The summaries consist of topical terms that characterises each cluster and a number of typical titles that sample the contents of the cluster. The user may select the summaries forming a sub- selection for iterative examination. The clustering and re- clustering is done so that different topics are seen depending on the sub-collection cluster. The schematic diagram of Scatter/Gather clustering algorithm is shown in figure 3. Scatter/Gather may be applied to the entire corpus, in which case static off-line computations may be exploited to speed dynamic online clustering. The use of Scatter/Gather successfully conveys some of the content and structure of the corpus. However, Scatter/Gather is less effective than a standard similarity search when the subjects are provided with a query.

C. Ranking of Clusters

A group of clusters are obtained after applying the document clustering algorithm each of which contains more or less relevant documents. By ranking the clusters we expect to determine reliable clusters and adjust the relevance score of documents in each ranked list such that relevant scores become more reasonable [28]. In order to find the ranked cluster of the query three measures are applied: Normalised Match Ratio (NMR), Normalised Order Ratio (NOR) and Log Odds Ratio (LOR).

NMR shows the number of terms related to the topic. It calculates the number of terms with high relevance with a given query and uses this to decide which cluster will be used. After which it shows them according to their rank.

a) Labelling-then-clustering approach*:* The labelling- then-clustering approach first identifies sets of documents that share phrases and extracts these phrases as candidate cluster labels. Candidate cluster labels are ranked and some of them are selected as the final cluster labels. Base clusters are created according to these cluster labels. Grouper [14] adopts a phrase-analysis algorithm called Suffix Tree Clustering STC, in which snippets sharing the same sequence of words are grouped together. Suffix Tree Clustering STC is a linear time clustering algorithm that is based on identifying the phrases that are common to groups of documents. A phrase in our context is an ordered sequence of one or more words. We define a base cluster to be a set of documents that share a common phrase. STC has three logical steps: (1) document cleaning (2) identifying base clusters using a suffix tree, and (3) combining these base clusters into clusters as shown in Figure 4. A Suffix tree is a

b) data structure that admits efficient string matching and

c) querying. Suffix trees have been studied and used extensively in fundamental string problems such as large volumes of biological sequence data searching [14], approximate string matches [5] and text features extraction in spam email classification [17]. In suffix tree document model, a document is considered as a string consisting of words, not characters. In Zamir and Etzioni STC algorithm, after the suffix tree construction, the overlap of the different clusters is calculated and the clusters are merged if they have more than 50% overlap. The merging method is fast but it neglects the similarity between the known overlapping parts. Another problem in the merging algorithm is that it can lead to too many clusters in hundreds and thousands with only a small amount of documents in each of it frustrating the browser to locate the desired information. Further, a new cluster merging algorithm of suffix tree clustering introduces the well known cosine similarity algorithm into the cluster merging process

V. CONCLUSION

We have mentioned that clustering is used to improve the IR from the collection of documents. The need of Information Retrieval mechanism can only be supported if the document collection is organised into a meaningful structure, which allows part or all the document collection to be browsed at each stage of a search. This has prompted researchers to re- examine the process of cluster based information retrieval. In the process of IR, we can calculate similarity coefficients between the query and the documents and search which clusters of documents best correspond to the query. This way of calculation is less time consuming for searching documents with high similarity than calculation of similarity coefficients between the query and individual documents. Numerous studies and anecdotal evidence hint that document clustering can be a better way of organising the retrieval results. Hierarchical algorithms start with established clusters, and then create new clusters based upon the relationships of the data within the set. All the relationships are analysed in the hierarchical algorithms which tend to be costly in terms of time and processing power. Moreover agglomerative hierarchical clustering does not do well because of the nature of documents, i.e., nearest neighbours of documents often belong to different classes.

REFERENCES

[1] W. B. Croft, “Organising and Searching Large Files of Documents”. PhD thesis, University of Cambridge, October (1978).

[2] C.J.van Rijsbergen, “Information Retrieval”, Butterworths, London, 2nd ed.(1979).

[3] W. B. Croft and R. H. Thompson, (I3R: A new approach to the design of document retrieval systems”, Journal of the American Society for Information Science, pp.389- 404, (1987).

[4] P. Willett, Recent, “Trends in Hierarchical Document Clustering”, A Critical Review, Information Processing & Management Vol. 24, No. 5, (1988).

[5] Anil K. Jain, Richard C. Dubes, Algorithms for clustering, Prentice Hall (1988).

[6] R.B.Allen, P.Obry and M.Littman, “An interface for navigating clustered document set returned by queries”, In Proceddings of ACM Conference on Organisational Computing Systems, pp.166-171, (1993).

[7] D. R. Cutting, D. R. Karger, and J. O. Pedersen, “Constant interaction-time Scatter/Gather browsing of very large document collections”, In Proceedings of ACM SIGIR, pp.126-134, (1993).

[8] D. Dubin, “Document analysis for visualization”, In Proceedings of ACM SIGIR, pp.199-204, July (1995).

[9] M. A. Hearst and J. O. Pedersen, “Reexamining the cluster hypothesis: Scatter/gather on retrieval results”. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 76-84, (1996).

[10] R. Weiss, et al., “HyPursuit: a hierarchical network search engine that exploits content-link hypertext clustering”, Seventh ACM Conference on Hypertext, pp. 180–193, (1996).

[11] J.Allan, “Building hypertext using information retrieval”, Information Processing and Management, pp.145-159, (1997).

[12] J. Allan, J. Callan, B. Croft, L. Ballesteros, J. Broglio, J. Xu, and H. Shu. “Inquery at TREC-5”. In Fifth Text REtrieval Conference (TREC-5), pp. 119-132,(1997).

[13] J. Pitkow, P. Pirolli, “Life, death, and lawfulness on the electronic frontier”, Proceedings of ACM CHI'97, pp. 383–390, (1997).

[14] O. Zamir and O. Etzioni, “Grouper: A dynamic clustering interface to web search results”. In Proceedings of the 8th International World Wide Web Conference, Toronto, Canada, May (1999).

[15] O. Zamir and O. Etzioni, “Web Document Clustering: A Feasibility Demonstration”, In Proceedings of the 19th International ACM SIGIR Conference on Research and Development of Information Retrieval (SIGIR’98),(1998).

[16] A. Leuski and J. Allan, “Evaluating a visual navigation system for a digital library”. International Journal on Digital Libraries, pp.170-184, (2000).

[17] C.W. Wen, et al., “A distributed hierarchical clustering system for web mining, International Conference on Web- Age Information Management (WAIM2001)”, pp. 103– 113, (2001).

[18] Y.Wang, M. Kitsuregawa, “Use link-based clustering to improve web search results, International Conference on Web Information Systems Engineering (WISE)”, pp.119– 128, (2001).

[19] Meyer zu Eissen and Stein, “Analysis of Clustering Algorithms for Web-based Search”, Paderborn University, pp.168-178, (2002).

[20] H. Zeng, Q. He, Z. Chen, and W. Ma, “Learning to cluster web search results”, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (2004).

[21] S. Osinski, J. Stefanowski and D. Weiss, Lingo: “Search results clustering algorithm based on singular value decomposition”, In Proceedings of the International Conference on Intelligent Information Systems (IIPWM’04), Zakopane, Poland, (2004).

[22] J.Han and M.Kamber, “DataMining: concepts and techniques (chapter9)”. Intelligent Data base Systems Research Lab, School of Computing Science, Simon Fraser University, Canada, (2005).

[23] Yue Xu, Hybrid, “Clustering with Application to Web Mining”, (2005).

[24] S. Sambasivam and N.Theodosopoulus, “Advanced Data Clustering Methods of Mining Web Documents”, Information Science and Information Technology, Vol 3, pp.563-578, (2006).

[25] J.W.R. Li, “A New Clustering Merging Algorithm of Suffix Tree Clustering”, In IFIP International Federation for Information Processing, Intelligent Information Processing III, eds. Z.Shi, Shimohara K, Feng D, Boston:Springer, vol.228, pp.197-203, (2006).