INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND CREATIVE ENGINEERING (ISSN:2045-8711)
VOL.7 NO.08 AUGUST 2017, IMPACT FACTOR: 1.04

A Survey on Document Clustering Using Wordnet

M.Sangeetha
Assistant Professor, PG and Research Department of Computer Science
Kaamadhenu Arts and Science College, Sathyamangalam, Tamil Nadu, India.
S.Subasri & T.Priyanka
M.Phil Research Scholar, PG and Research Department of Computer Science,
Kaamadhenu Arts and Science College, Sathyamangalam, Tamil Nadu, India.

Abstract- WordNet is connected to several databases of the semantic web. WordNet is also commonly re-used via mapping between the WordNet synsets and the categories from ontologies. Most often, only the top-level categories of wordnet are mapped. It is used for a number of different purpose in information systems, including word-sense disambiguation information retrieval, automatic text classification, automatic text summarization, machine translation and even automatic crossword puzzle generation. Mostly this information data is stored in unstructured text. This large data developed has lead to the need of its systematic clustering for easy data retrieval organization and summarization, typically called as data mining. In this paper we Present document clustering using wordnet used different attributes and algorithm. Wordnet based algorithm is used for semantic similarity measure. It is designed to solve problems in text clustering. Semantic algorithm is compared with using all algorithm, Which proved to be more efficient and provides more pure clusters.

Keywords- Suffix Tree, Lingo, Suffix Array, Information Retrieval, Search Engine, Semantic, Tree Clustering, Document Clustering.

1. INTRODUCTION

Information Retrieval plays an important role in our daily life and its largest role is observed in search engines. Most users rely on Web search engines to look for specific information from the Web. These search engines often return a long list of search results that would be ranked by their relevance to the given query. Web users have to go through the long list and inspect the titles, and snippets sequentially to recognize the required results. Filtering the search engines' results consumes the users effort and time especially when multiple sub-topics of the given query are mixed together[1][2]. This paper describes how to overcome some of the major limitations in the current search engines. We proposed a multi-agent based information retrieval system to enhance the search process. We used different types of agents each of them has its own responsibility. We organize the results of Web search engine by clustering them into different categories for a given query. We utilized WordNet ontology and several approaches to cluster results in appropriate category according to WordNet synsets.

Search engines are an invaluable tool to retrieve information from the internet. On the other hand they tend to return an enormous amount of search results and this causes a time consuming task to find the relevant ones. Moreover, if the relevant results do not occur in the first part of the returned results, then the user may fail to find them[4][5]. Nowadays, the development of a search results clustering system involves semantic search results clustering, which in turn uses the semantic meaning of words to cluster. This idea considers semantically related words such as synonyms or hyponyms for increasing the quality of clusters. In 2010, Ahmed Sameh and Amar Kadray. proposed the Semantic Lingo algorithm which uses synonyms to extract phrase terms to use as discovery cluster labels.

The Semantic Lingo algorithm extends the Lingo techniques by adding semantic recognition, particularly using the WordNet database to achieve semantic recognition. Semantic recognition can improve the quality of the clusters generated. Suffix tree clustering, as a fast, incremental, linear time clustering algorithm, has been widely concerned about. Carrot2, a well-known open source clustering search engine has an implementation based on suffix tree clustering algorithm[6][7][8]. However, label-contained and duplication exist in the results of the clustering. It's an added burden for users to get their interest information.

This paper is aiming to solve this problem. The existing search engines always come out with a long list of results for the given query and they are ranked by their relevance to the same query. Information retrieval and ranking functions are vital to the search engines[8][9]. To address the above challenges, some effective approaches such as web pages categorizing (Yahoo!), query classification and search results clustering have been used or proposed. And search results clustering have been proved to be a more effective way, which is an automatic, online grouping similar documents in the search results returned by a search engine into a hierarchy of labeled clusters.

Based on the model described by the third approach, we argue that there are four key factors for search results clustering as The quality of cluster labels. Having a meaningful, unambiguous label for the text clustering is very important[11][12]. The accuracy of clustering results. Documents in the same cluster should have the consistent theme. A relatively short response time. Moreover, if the relevant results do not occur in the first part of the returned results, then the user may fail to find them. A possible solution to this problem is use of the search results clustering

techniques that works on snippet– a short text summarizing the context of search results. Search results clustering engine's main role is to cluster search results into different groups of relevant data and create a navigator to easily access the relevant search results for users. The development of a search results clustering system involves semantic search results clustering, which in turn uses the semantic meaning of words to cluster. This idea considers semantically related words such as synonyms or hyponyms for increasing the quality of clusters[13][14][15].

2. RELATED WORKS

A) Semantic Clustering Approach Based Multi-agent System for Information Retrieval on Web

Document clustering is an important technology which helps users to organize the large amount of online information, especially after the rapid growth of the Web. This paper focuses on semantic document clustering method and its application in search engine. We proposed a multi-agent based information retrieval system to enhance the search process. The agents retrieve the results of Web search engine and organize the results by clustering them into different categories for a given query. We utilized WordNet ontology and several approaches to cluster results in appropriate category according to WordNet synsets. The experiment shows that semantic clustering work better than original clustering.

In this paper, we investigated the problem of how to cluster the search result from search engines. Queries are often ambiguous because many words have multiple meanings. [1][2]By clustering the search results based on the semantic of the query term, it makes it easier for users to identify relevant results from the retrieved results. We proposed a modified version of the lingo algorithm that combines both WordNet ontology and clustering techniques. Our preliminary experimental results indicated that our semantic clustering algorithm is effective, achieving an accuracy of about 90%. We also showed that this algorithm is significantly better than original lingo cluster by about 6.39%. We plan to continue this research in the following directions[3]. First, we will work on some criteria to avoid clusters overlapping that mean document cannot be assigned to more than one cluster. Second, we will try to remove the near duplicate cluster label.

B) Semantic Suffix Tree Clustering

This paper proposes a new algorithm, called Semantic Suffix Tree Clustering (SSTC), to cluster web search results containing semantic similarities. The distinctive methodology of the SSTC algorithm is that it simultaneously constructs the semantic suffix tree through an on-depth and on-breadth pass by using semantic similarity and string matching. The semantic similarity is derived from the WordNet lexical database for the English language[4][5]. SSTC uses only subject-verb-object classification to generate clusters and readable labels. The algorithm also implements directed pruning to reduce the sub-tree sizes and to separate semantic clusters. Experimental results show that the proposed algorithm has better performance than conventional Suffix Tree Clustering (STC).

This paper proposes a new algorithm, called Semantic Suffix Tree Clustering (SSTC), that uses the meaning of the words to cluster. SSTC can cluster documents that share a semantic similarity. Specific cluster are returned in a readable form. Additionally, the SSTC can improve the performance of approaches that use the original STC algorithm because it can cluster semantically similar documents, reduce the number of nodes and reach higher precision. For future work we plan to extend the SSTC algorithm to hierarchical clustering.

C) Semantic-based Hierarchicalize the Result of Suffix Tree Clustering

Suffix tree clustering is a fast, incremental, linear time clustering algorithm, but there are synonymous and label-contained relations among the result clusters. So just return these results to the users directly, would give them an added burden. In response to this problem, this paper presents a method that merging the semantic duplicate clusters and hierarchicalizing the label-contained clusters. The experimental results show that this method can effectively remove semantic duplication and hierarchicalize label-contained clusters clearly[6][7]. It improves the organization of clustering results. To the STC search engine, this will provide users with better results and better classification.

In this paper, through semantic-based processing the results of the STC, merging synonyms clusters, hierarchicalizing label-contained clusters, improve the organization of clustering results. To the STC search engine, this will provide users with better results and better classification. This can help users both in locating interesting documents more easily and in getting more clearly overview of the retrieved document set[8]. In this paper, the semantic-based hierarchicalizing process only uses synonymy relationship. How to make use of antonym and other important semantic relations to further improve the organization of the clustering results is a future work.

D) A Relative Study on Search Results Clustering Algorithms - K-means, Suffix Tree and LINGO

The performance of the web search engines could be improved by properly clustering the search result documents. Most of the users are not able to give the appropriate query to get what exactly they wanted to retrieve. So the search engine will retrieve a massive list of data , which are ranked by the page rank algorithm or relevancy algorithm or human judgment algorithm. The user will always find himself with the unrelated information related to the search due to the ambiguity in the query by the user[9][10]. Evaluating the performance of a clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm In this paper a comparative analysis is done on three common search results of clustering algorithms to study the performance enhancement in the web search engine. If we effectively organize the web documents through the proper means of clustering techniques, we could definitely increase the performance of the search engines.

A systematic evaluation of the three clustering algorithms viz., Suffix tree clustering Lingo, and K-Means using multiple test collections and evaluation measures . It turns out that STC works well, when one wants to get a quick overview of documents relevant to distinct subtopics, whereas clustering is more useful when one is interested in retrieving multiple documents relevant to each subtopic. Each algorithm has its

own merits and demerits, Lingo produces high cluster diversity, the Small outliers are highlighted well, In STC and K-means algorithms the small outliers are rarely highlighted [11].In Lingo the number of clusters produced are more when compared to other two algorithms..With respect to the cluster labels , in LINGO they are descriptive but lengthy , not very descriptive in K-Means ,but in STC cluster labels are small but very appropriate. The Scalability is high in STC compared to Lingo and K-Means. Other features of K-Means clustering are Running time: O(KN) (K = number of clusters) Fixed threshold ,Order dependent. Features of STC are Overlapping clusters, Non-exhaustive Linear time, and High precision.

E) Search Results Clustering Based on Suffix Array and VSM

With the rapid growth of web pages, search engines will usually present a long ranked list of documents. The users must sift through the list with "title" and "snippet" (a short description of the document) to find the desired document. This method may be good for some simple and specific tasks but less effective and efficient for ambiguous queries such as "apple" or "jaguar". To improve the effect and efficiency of information retrieval, an alternative method is to automatically organize retrieval results into clusters[12][13]. This paper presents an improved Lingo algorithm named Suffix Array Similarity Clustering (SASC) for clustering web search results. This method creates the clusters by adopting improved suffix array, which ignores the redundant suffixes, and computing document similarity based on the title and short document snippets returned by Web search engines. Experiments show that the SASC algorithm has not only a better performance in time-consuming than Lingo but also in cluster description quality and precision than Suffix Tree Clustering.

In the paper, we propose a search results clustering algorithm named SASC. And its main contributions are the efficiency of suffix array is improved by ignoring the redundant suffixes. We also proved that the equivalent cluster results can be obtained by analyzing the matrix of consequence rather than by computing SVD[14][15]. This method takes far less time than Lingo. Furthermore, SASC supports hierarchical structure. In the future, we intend to further improve the time efficiency and the accuracy as well as consider other information such as the user's interaction with the clustering results for adaptive clustering.

F) Clustering of Web Search Results Using Semantic

Clustering is related to data mining for information retrieval. Relevant information is retrieved quickly while doing the clustering of documents. It organizes the documents into groups; each group contains the documents of similar type content. Different clustering algorithms are used for clustering the documents such as partitioned clustering (K-means Clustering) and Hierarchical Clustering (Agglomerative Hierarchical Clustering (AHC)). This paper presents analysis of Semantic Suffix Tree Clustering (SSTC) Algorithm and other clustering techniques (K-means, AHC, and Lingo). SSTC perform the clustering and make the clusters based on synonyms shared between the documents. SSTC is faster clustering algorithm for document clustering as it is incremental.

The paper presents the analysis of different clustering techniques such as partitioned clustering and hierarchical clustering. K-means presents the Partitioned clustering and Agglomerative Hierarchical Clustering presents the Hierarchical clustering. Also it analyses Semantic Lingo Algorithm. It introduces an algorithm for web search result clustering known as Semantic Suffix Tree Clustering Algorithm[16][17][18]. The paper proposes the main steps as to identify base clusters, merging the base clusters. SSTC can improve the performance of approaches that use the original STC algorithm because it can cluster semantically similar documents, reduce the number of nodes and reach higher precision.

3. ANALYSIS AND DISCUSSION

In this section we conduct several experiment to validate the effectiveness of the proposed approach. The process was various clustering algorithm to the document collections and compared their precision. This method takes far less time that Lingo, furthermore SASC supports hierarchical structure. We intend to further improve the time efficiency and the accuracy as well as consider other information such as the user's interaction with the clustering results for adaptive clustering. Semantic algorithm is compared with Lingo, Which proved to be more efficient and provides more pure clusters. Semantic Lingo increases efficiency and provide more relevant result. The higher number of matrix transformation leads to demanding memory requirements. So Semantic algorithm is designed for specific application like web search result clustering.

Author	Algorithm	Attributes	Results
Bassma S, Alsulami, Maysoon F, Abulkhair Fathy A Essa	Clustering Algorithm	WordNet Semantic clustering	The preliminary experimental results indicated that our semantic clustering algorithm is effective, achieving an accuracy of about 90%. We also showed that this algorithm is significantly better than original lingo cluster by about 6.39%.
Jongkol Janruang, Sumanta Guha	SSTC STC	Semantic Search Result Clustering Text Clustering	The SSTC can improve the performance of approaches that use the original STC algorithm because it can cluster semantically similar documents, reduce the number of nodes and reach higher precision.
Guodong Hu, Wanli Zuo, Fengling He, Ying Wang	Suffix Tree Clustering Grouper	Semantic Hierarchical Suffix Tree Clustering	The STC search engine, this will provide users with better results and better classification. This can help users both in locating interesting documents more easily and in getting more clearly overview of the retrieved document set. In this paper, the semantic-based hierarchicalizing process only synonymy relationship.
Mahalakshmi R, Lakshmi Prabha V	Suffix Tree Clustering K Means Clustering Algorithm LINGO	Information Retrieval Search Engines	The Scalability is high in STC compared to Lingo and K-Means. Other features of K-Means clustering are Running time: O(KN) (K = number of clusters),Fixed threshold, Order dependent. Features of STC are Overlapping clusters, Non-exhaustive, Linear time and High precision.

Shunlai Bai, Wenhao Zhu, Bofeng Zhang	Search Result Clustering Algorithm named Suffix Array Similarity Clustering	Suffix Array Suffix Tree Lingo	We propose a search results clustering algorithm named SASC and The efficiency of suffix array is improved by ignoring the redundant suffixes. We also proved that the equivalent cluster results can be obtained by analyzing the matrix of consequence rather than by computing SVD.
Shelke p p, Dhopte S V, Alvin A S	SSTC STC	Document Clustering Partitioned Clustering	It introduces an algorithm for web search result clustering known as Semantic Suffix Tree Clustering Algorithm. The paper proposes the main steps as to identify base clusters, merging the base clusters.

6. CONCLUSION

In this paper proposed a novel method ,document clustering algorithm which identifies key concept and automatically generates Ontology's for users to conceptualize document corpora. It presents a novel approach termed as document clustering using algorithms based on concept of the text data. This concept driven approach is executed on Wordnet. It introduces an algorithm for web search result clustering known as semantic suffix tree clustering algorithm. On internet is a significant and challenging problem. Several new concepts and the mining problem are formally defined and a group of algorithm are designed and combined to systematically solve this problem. As compare different web search clustering algorithm like STC and SHOC does not reduce the high dimension of the text document hence its complexity is quite high for large text data based which ignores the semantic and lexical relationship between words, proposed new algorithm called Lingo Semantic for clustering.

REFERENCES

[1] Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y., Ma, J.: "Learning to cluster Web search results. In: SIGIR '04". Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY,USA, ACM Press (2004) 210–217

[2] T. de Simone and D. Kazakov. "Using WordNet Similarity and Antonymy Relations to Aid Document Retrieval". Recent Advances in Natural Language Processing (RANLP), 2005.

[3] M. A. Hearst, J. O. Pedersen. "Re-examining the Cluster Hypothesis: Scatter/Gather on Retrieval Results". In Proceedings of the ACM SIGIR Conference, 1996.

[4] Ahmed, M.S., Amar, M.K.: Semantic Web Search Results Clustering Using Lingo And Wordnet. In: IJRRCS: Kohat University of Science and Technology (KUST), Vol. 1, No 2, pp. 71–76. , Pakistan (2010)

[5] Stanislaw, O., Jerzy, S.,: An algorithm for clustering of web search results. Master Thesis, Poznan University of Technology, Poland, June 2003

[6] Carpineto, C., Osinski, S., Romano, G., and Weiss, D.: A survey of Web clustering engines. In: ACM Computing Surveys, Volume 41 , Issue 3, pp. 1–38. ACM, New York, USA (2009)

[7] E. Ukkonen, "On-line construction of suffix trees," Algorithmica, vol. 14, 1995, pp 249-260.

[8] Oren Zamir ,Oren EtzioniO. "Grouper: A Dynamic Clustering Interface to Web Search Results," University of Washington. Department of Computer Science and Engineering. 1999K. Elissa

[9] O. Zamir and O. Etzioni, "Web document clustering: a feasibility demonstration," in: Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98),1998, pp 46-54.

[10] C Carpineto, and S Osiński, G Romano, and D Weiss, "A survey of web clustering engines", ACM Computing Surveys (CSUR), ACM, 2009.

[11] H Cao, DH Hu, D Shen, D Jiang, JT Sun, E Chen, and Q Yang, "Context-Aware query classification", Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, Boston, USA, 2009, pp. 3-10.

[12] Oren Zamir, and Oren Etzioni, "Web Document Clustering: A Feasibility Demonstration", Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, Melbourne, Australia, 1998, pp. 46-54.

[13] Pushpalatha, Ram Chatterjee: An Analytical Assessment on Document Clustering. In: I. J. Computer Network and Information Security, 2012, 5, 63-71 published Online June 2012

[14] Ahmed, M.S., Amar, M.K.: Semantic Web Search Results Clustering Using Lingo and Wordnet. In: IJRRCS: Kohat University of Science and Technology (KUST), Vol. 1, No 2, pp. 71–76. , Pakistan (2010)

[15] Stanislaw, O., Jerzy, S.,: An algorithm for clustering of web search results. Master Thesis, Poznan University of Technology, Poland, June 2003

[16] B. R. Prakash and M. Hanumanthappa, "Web Snippet Clustering and Labeling using Lingo Algorithm", International Journal of Advanced Research in Computer Science, vol. 3, no. 2, pp. 262-265, 2012

[17] Carpineto, S. Osinski, G. Romano, D. Weiss. A Survey of Web Clustering Engines. ACM Computing Surveys (CSUR), 41(3): Article 17, 2009

[18] Tingting Wei, Yonghe Lu, Huiyou Chnag, Qiang Zhou, Xianyu Bao ," Sematic approach for text clustering using WordNet and lexical chains", Expert Systems with Applications,Volume 42, Issue 4, March 2015, pp. 2264–2275.