OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

A Review Paper on Description Oriented Methods for Clustering Of Web Search Result

Poonam Churi[1], Sujata Kolhe[2]
**(Datta Meghe college of engineering, Airoli, Navi Mumbai, India)*
*** (Datta Meghe college of engineering, Airoli, Navi Mumbai, India)*

ABSTRACT:- *In the age of increasing information availability, many techniques, such as Web document clustering and information visualization, have been developed to ease understanding of information for users. However, most of these methods do not help users directly understanding Key concepts and their semantic Relationship in document corpora, which are critical for capturing their conceptual structures Labelling Cluster is crucial because meaningless or confusing labels may mislead users to check wrong clusters for the query and lose extra time. Additionally, labels should reflect the contents of documents within the cluster accurately. To be able to label clusters effectively, a new cluster labelling methods are introduced. More emphases was given to/produce compréhensible and accurate cluster labels in addition to the discovery of document clusters. The main goal of this paper is to discuss various web document search results methods We adopt a comparative evaluation strategy to derive the relative performance of the methods with respect to the four prominent search result clustering methods: Suffix Tree Clustering, Semantic Hierachical Online Clustering, Lingo and Cuckoo.*

Keywords: *Document Clustering, Information Retrieval, Relevant, Web Search Result.*

I. INTRODUCTION

The question of how to find information of interest in the Internet is raised by the Web Search Problem. Find the set of documents on the Web relevant to a given user query. The definition differs from the well-known Information Retrieval Problem: given a set of documents and a query, determine the subset of documents relevant to the query, in several aspects. First of all, it recognizes the fact that the Web is a highly dynamic collection of documents, which makes many of the classic indexing schemes unsuitable. Secondly, due to the sheer size of the Internet, the Web Search Problem assumes that only a limited number of documents will actually be matched against the input query and even a smaller proportion of them will be finally viewed by the user. Thus, special attention must be given to helping the user choose the most relevant documents first. [1]

Web Clustering Engines are the systems that perform clustering of web search results. This systems group the results returned by a search engine into a hierarchy of labeled clusters (also called categories). Several clustering engines for web search results have been implemented. Grouper employs a novel, phrase-based algorithm called Suffix Tree Clustering (STC),[2] Carrot employs SHOC (Semantic Hierarchical Online Clustering) [6] AND Lingo[4] algorithms for clustering of documents. Other examples of clustering engines can be the Scatter/ system, the Class Hierarchy Construction Algorithm and iBoogie. Vivisimo [uses an intriguing, yet not publicized, technique for organizing the search results into hierarchical and very well described thematic groups. this paper presents comparative assessment of widely used available clustering engines for fast retrieval of the data. Such systems usually consist of four main components: search results acquisition, preprocessing of input, cluster construction and labeling, and visualization of resulting clusters. Comparative study had been done on various web search clustering methods to show the future improvement in the field of document clustering. [5]

II. STEPS IN DOCUMENT CLUSTERING

A) *Search result acquisition* - The search results acquisition component begins with a query defined by the user. Based on this query, a document search is conducted in diverse data sources, in this case in the traditional web search engines such as Google, Yahoo! In general, web clustering engines work as Meta search engines and collect between 50 to 200 results from traditional search engines. These results contain as a minimum a URL, a snippet and a title. [1]

B) *Preprocessing-* the preprocessing of search results comes next. This component converts each of the search results (as snippets) into a sequence of words, phrases, strings or general attributes or characteristics, which are then used by the clustering algorithm. There are a number of tasks performed on the search results, including: removing special characters and accents, the conversion of the string to lowercase, removing stop words, stemming of the words and the control of terms or concepts allowed by a vocabulary.

C) *Cluster construction and labeling***-** Once preprocessing is finished, cluster construction and labeling is commenced, making use of three types of algorithm data-centric, description-aware and description-centric. Each of these builds clusters of documents and assigns a label to the groups. [3]

D) *Visualization* - Finally, in the visualization step, the system displays the results to the user in hierarchically organized folders. Each folder seeks to have a label or title that represents well the documents it contains and that is easily identified by the user. As such, the user simply scans the folders that are actually related to their specific needs. The presentation folder tree has been adopted by various systems such as Carrot2, Yippy, SnakeT, and KeySRC, because the folder metaphor is already familiar to computer users. Other systems such as Grokker and Kart004 use a different display scheme based on graphs.

III. DIFFERENT WEB SEARCH CLUSTERING ALGORITHMS.

1) Suffix Tree Clustering (STC)

The Suffix Tree Clustering (STC) [2] algorithm groups the input texts according to the identical phrases they share. The rationale behind such approach is that phrases, compared to single keywords, have greater descriptive power. This results from their ability to retain the relationships of proximity and order between words. A great advantage of STC is that phrases are used both to discover and to describe the resulting groups. The Suffix Tree Clustering algorithm works in two main phases: base cluster discovery phase and base cluster merging phase. In the first phase a generalized suffix tree of all texts' sentences is built using words as basic elements. After all sentences are processed, the tree nodes contain information about the documents in which particular phrases appear. Using that information documents that share the same phrase are grouped into base clusters of which only those are retained whose score exceeds a predefined Minimal Base Cluster Score. In the second phase of the algorithm, a graph representing relationships between the discovered base clusters is built based on their similarity and on the value of the Merge Threshold. Base clusters belonging to coherent sub graphs of that graph are merged into final clusters. A clear advantage of Suffix Tree Clustering [2] is that it uses phrases to provide concise and meaningful descriptions of groups. The method basically involves the use of a tree structure to represent shared suffixes between documents. Based on these shared suffixes, they identify base clusters of documents, which are then combined into final clusters based on a connected-component graph algorithm. The tree kind structure contains all suffix substrings of the document d. The internal node has at least two children. Each edge is labeled with a nonempty substring of a document called a phrase. Then, each leaf node in the suffix tree designates a suffix substring of a document; each internal node represents a common phrase shared by at least two suffix substrings. If the documents are more similar, they share more internal nodes. The original Suffix Tree Clustering (STC) algorithm is developed based on the Suffix Tree Document (STD) model. In detail, the STC algorithm has three logical steps. [2]

Step 1. The common suffix tree generation
Step 2. Base cluster selection.
Step 3. Cluster merging

2) Semantic Hierarchical Online Clustering (SHOC)

To overcome the STC's low quality phrases problem, in SHOC Zhang and Dong [6] introduce two novel concepts: complete phrases and a continuous cluster Definition. The SHOC algorithm works in three main phases: complete phrase discovery phase, base cluster discovery phase and cluster merging phase. In the first phase, suffix arrays are used to discover complete phrases and their frequencies in the input collection. In the second phase, using Singular Value Decomposition a set of orthogonal base clusters is obtained. Finally, in the last phase, base clusters are merged into a Hierarchical structure. One of the drawbacks of SHOC is that Zhang and Dong [6] provide only vague comments on the values of Merge Threshold 1 and 2 and the method which is used to describe the resulting clusters.

3) *Lingo-*

To overcome drawbacks consist by SHOC web search clustering algorithm we used new algorithm called Lingo. We have decided to use a slightly modified version of SHOC's complete phrase discovery algorithm. We also employ the Singular Value Decomposition, but rather than use it to find cluster contents, we utilize this technique to identify meaningful group labels [7].

Step 1: Data preprocessing -The usual stemming (Porter stemmer, Lametyzator), stop words making, text segmentation heuristic.

Step 2: Frequent phrase extraction and cluster label induction - Discover complete phrases in the input text maximum length term subsequences occurring at least term frequency threshold times do not cross sentence boundaries no stop words at ends. Identifies the abstract concepts that best describe the input snippet collection and uses frequent phrases to construct a human-readable representation of these concepts. And further these are matched with against a series of queries.[7]

Step 3: Cluster content allocation - is calculating group scores as a product of the label score and the number of snippets in the group.

4) *Cuckoo –*

The new algorithm, called Web Document Clustering based on the Cuckoo Search Algorithm [10] is a description-centric algorithm for the clustering of web results, which was inspired by the new meta-heuristic algorithm, Cuckoo Search. All Cuckoo Search optimization technique is introduced by Yang and Deb recently [10]. Cuckoos have a belligerent reproduction tactic that involves the female laying her fertilized eggs in the nest of another species so that the surrogate parents unwittingly raise her brood [10]. Sometimes the cuckoo's egg in the nest is revealed and the surrogate parents throw it out or dump the nest and start their own brood elsewhere. The cuckoo search optimization algorithm considered various design parameters and constraints, the three main idealized rules on which it is based are as follows1)Each cuckoo lays one egg at a time, and dumps its egg in randomly chosen nest;2) The best nests with high quality of eggs will carry over to the next generations 3) The number of available host nests is fixed, and the egg laid by a cuckoo is discovered by the host bird with a probability pa€ [0, 1]. In this case, the host bird can either throw the egg away or abandon the nest, and build a completely new nest Cuckoo Search Clustering Algorithm based on levy flight [8] is designed as a clustering algorithm from Cuckoo Search Optimization algorithm to locate the optimal centroids of the cluster. In web document clustering area, it is possible to view the clustering problem as an optimization problem that locates the optimal centroids of the clusters rather than an optimal partition finding problem. This algorithm aims to group a set of input samples (data points) into clusters with similar features. It will work without the knowledge of the class of the input data during the process. In this algorithm, new cuckoo [10]

Solutions will be moved by using levy flight. Steps for cuckoo algorithm as follows:

1. Begin
(Parameter Initialization- no of clusters, no of host nests)
2. Consider NH host nests containing 1 egg (solution) each
3. for each solution of host i
4.Initialize xi to contain k randomly selected cluster centroids (corresponding to k clusters), as xi = (mi,1,….,mi,j,…mi,k) where mi,k represents the kth cluster centroid vector of ith cluster centroid vector of ith host.
End for loop
5. for t iterations
6. for each solution of host i of the population
7. For each data document zp
8. Calculate distance d (Zp, mj, k) from all cluster centroids Ci, k by using Cosine Similarity Distance eq-2
9. Assign zp to Ci, k by
d(zp,mj,k) = mink= 1…k { d(zp,mj,k) }
End for loop in step 7
10. Calculate fitness function f (xi) for each host nest i by eq-3
11. End for loop in step 6
12. Replace all the nests except for the best one by new Cuckoo eggs produced with levy flight from their positions
13. A fraction pa of worse nests are abandoned and new ones are built randomly
14. Keep the best solutions (or nests with quality solutions)
15. Find the current best solution

End for loop in step 5
16. Consider the clustering solution represented by the best solution
17. End.

IV. COMPARISION BETWEEN WEB CLUSTERING METHODS

Table 1.Label evaluation

Algorithms	Comprehensibility	Descriptiveness	Uniqueness	Non Redundancy
Suffix Tree Clustering	0.70	0.73	0.84	0.90
Suffix Hierarchical Online Clustering	0.78	0.80	0.93	0.98
Lingo	0.80	0.93	0.89	0.93
Cuckoo	0.83	0.98	0.94	0.93

V. CONCLUSION

In this paper we have done a survey on different types of web search result clustering methods and compare five types of methods to proper label evaluation. STC is an incremental and liner time (in the document collection size) algorithm, which creates clusters based on phases shared between documents .STC doesn't use VSM (Vector Space Model) to express text as bag-of-words, which is different from other clustering algorithms .STC is faster and more flexible than other standard snippets clustering methods. However, STC is still not perfect. It ignores the semantic information in snippets and in high-cost of space when the number of snippets is huge. There are two classic improved algorithms based on STC, SHOC and Lingo SHOC uses singular value decomposition (SVD) to discover the semantic information in term-document matrix generated by VSM. Lingo uses common phase discovery and LSI (Latent Semantic Indexing) to group snippets into meaningful clusters. Unfortunately, because their processes are still based on VSM, the semantic relationship between the words is not recognized explicitly. What's more, they are high-dimensional when applied to large numbers of snippets, which goes against the high space cost of STC. Cuckoo also improves clustering and label quality .main aim of such type of methods is to reduce overlapping rate and increases relevancy.

REFERENCES

[1]. Web Document Clustering: A Feasibility Demonstration" Oren Zamir and Oren Etzioni Department of Computer Science and Engineering University of Washington Seattle, WA 98195-2350 U.S.A.

[2]. The Suffix Tree Document Model Revisited Sven Meyer zu Eissen (Paderborn University, Germany smze@upb.de) Benno Stein (Bauhaus University Weimar, Germany benno.stein@medien.uniweimar.

[3]. Martin Potthast (Paderborn University, Germany beebop@upb.de A Search Result Clustering Method using informatively Named Entities" Hiroyuki Toda NTT Cyber

[4]. Solutions Laboratories, NTT Corporation WIDM'05, November 5, 2005, Bremen, Germany. Copyright 2005 ACM 1-59593-194-5/05/0011 ...$5.00 S.Osinski and D. Weiss. A concept-driven algorithm for clustering search results. 20(3):48–54, 2005.

[5]. M. Steinbach, et al., "A comparison of document clustering techniques," in KDD workshop on text mining, Boston, MA, USA., 2000, pp. 1-20

[6]. 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, Pages 2264–2275.

[7]. Stanislaw Osiríski and Dawid Weiss. Conceptual clustering using Lingo algorithm: Evaluation on Open Directory Project data. Submitted to Intelligent Information Systems Conference 2004, Zakopane, Poland, 2003

[8]. Yang, X.S., Deb, S.: Cuckoo search via Levy flights. In: Proc. of the World Congress on Nature and Biologically Inspired Computing, India, pp. 210–214 (2009)

[9]. Andrea Tagarelli, George Karypis," A segment-based approach to clustering multi-topic documents", Knowl INF Syst (2013) Springer-Verlag London Limited 2012.

[10]. Moe Moe Zaw and Ei Ei Mon,Web Document Clustering Using Cuckoo Search Clustering Algorithm based on Levy Flight, International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 4 No. 1 Sep. 2013