Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 216 – 222

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Labeling of Web Search Result Clusters using Heuristic Search and Frequent Itemset

Mansaf Alam[a] , Kishwar Sadaf[a,*]

[a] *Jamia Millia Islamia,Jamia Nagar, New Delhi-110025, India*

Abstract

Clustering of search result is undoubtedly a tool that can provide the summarization of the millions of documents in a way where a user can easily locate his/her information. To guide user to the right cluster of documents, cluster labels should be meaningful and correctly representing the clusters. However significant a cluster is, if the label is not proper, user will never select it. In this paper, we present a method to label clusters based on their linking information. Our cluster labeling method is independent of any clustering method but restricted to only search result documents. We use heuristic search method to find all the linked documents of a cluster. If all or some documents of a cluster share hyperlinks, then we deduce label from these linked documents' titles using famous Apriori algorithm for frequent itemset mining. This removes the requirement of reviewing other members of a cluster in labeling process.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Web search result clustering; cluster labeling; heuristic search.

1. Introduction

The amount of information on the web is increasing day by day. Moreover this information is heterogeneous in nature. There is no uniformity. Anyone can upload data on the web in any format they want. To be among top in the

* Corresponding author. Tel.: +91-9911950408
E-mail address: kishwarsadaf@gmail.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.014

ranking of search engine result, people use different kind of techniques, without having any relevance to the query. To fetch relevant information from this web is a tricky business. Search engines return millions of pages or documents in an answer to a query. The result becomes even larger if the query is ambiguous as search engines try to retrieve documents for all the possible meaning of a query. Clustering of search result is a way to summarize this large amount of documents in form of groups where group members share similar qualities. There are many clustering engines available like Kartoo, Carrot2, Vivisimo etc. As search result clustering is being widely researched, research on cluster labeling is also going hand in hand [1,2,3,4]. Labeling of clusters is equally important. If the label of an accurately composed cluster does not define its contents, then there is a less chance that a user may select it even if it contains his desired information. Authors present a method to cluster documents based on key term labeling provided by the user[5]. Selecting terms for labeling is another thrust area of research. Mutual information, $\chi 2$ test, frequent itemset mining are some statistical methods which have been used in labeling. Labels can be deduced from finding the most important sentences and headlines in a news cluster[6]. Authors utilize the whole cluster to mine frequent words and differentiate the clusters on these frequent words[7]. Using external knowledge bases like Wikipedia, wordnet or ontologies is not new in the field of clustering. They are being employed in the labeling process also. Authors present a method to label clusters using Wikipedia[8]. In the paper[9], authors proposed a method which relies on the linked resources available on the web to infer labels. These link information are in the forms of parent-child and synonym links. Fumiyo et al[10] proposed a labeling method based on the assumption that all the salient words of a cluster share same hypernym. Authors propose a labeling or description defining method which is based on the SVM model[11]. Chen et al[12] , assert that their method of cluster labeling, ClusterMap, is well suited to entire dataset.

We propose a labeling method for web search result clusters i.e. our method is applicable to clusters produced from web search results. Web search results have many great features. One such feature is that some prominent pages related to a category share links. These linked pages' titles are searched for frequent itemset. Title of a page contains salient words which reflect the content of the page. We search for the frequent itemsets in a set of titles. The novelty of our approach is the use of heuristic search in combination with frequent itemset mining. Our experimental result shows that linked pages which participate in labeling process, produce good labels.

The rest of the paper is arranged as follows. Section 2 provides the work related to the labeling of web search result. Proposed method is described in section 3. Section 4 presents the experimentation and results followed by conclusion in section 5. Pages and documents are used interchangeably throughout the paper.

2. Related Work

Labeling of web search result cluster is an ongoing field of research. Authors propose a method where search result clustering is performed using cover coefficient and sequential k-means method[13]. The clusters are then labeled using term weighting. Filippo et al[14] , presents a method that clusters search result using k-means. The labeling of the clusters is done by using information gain (IG) on the terms of the cluster. Labeling can be done by selecting a candidate term from the clusters using association mining[15].

3. Proposed Method

We propose a method to label clusters of search result returned by a search engine. Label forming words are inferred from titles of documents of a cluster that shares hyperlinks. It is not necessary that all the pages of a cluster are linked. We hold that pages that are linked in the web search result clusters truly define them. For example for the query "puma", search engine Google returns thousands of pages. The result contains pages about Puma merchandiser, puma lion, puma web server, puma chocolate music band and many more other categories. We found that some pages of these categories are hyperlinked and accurately define their categories.

To find all the connected pages of a cluster, we apply our heuristic search method. Each page of a cluster contains hyperlinks to other pages. Our heuristic says that clusters tend to form where there are many links. For each page in a cluster, our method maintains a list of pages, which it connects. Pages in the list are also from the same cluster. After each and every page's list of connected pages are initialized, we apply our heuristic. In each list, we find the promising page, which has the corresponding largest list, and add its connected pages to it.

Input: A set of similar web page clusters, $C = \{c_1, c_2, ..., c_n\}$

Output: Cluster Labels

Step 1: Page Collection phase

For each $p_i = c \forall 1 \le i \le n$

Initially, $p_i = h$, (promising page)

Initialize each page's list of connected pages

$$l(h_i) = \cup p_j \mid i \ne j \;\&\; \forall 1 \le j \le n$$

Search for promising page h using heuristic

Expand p_i's list L of connected pages

$$l(a_i) \leftarrow l(h_i)$$

Filter $l(p_i)$ for new promising page

$$h(p_i) = p_j : \arg\max \left\| l(p_j) \right\| \forall 1 \le j \le n \;\&\; j \ne i$$

Update the list $l(p_i)$

Merge all lists into L

$$if \left((p_i \in l_m) \cap (p_{i'} \in l_{m'}) \right) \& \; l_m \ne l_{m'}$$

Merge $l_m \; and \; l_{m'}$

Step 2: Extract titles from pages in L

$$t_i = title(p_i \in L)$$

Algorithm for Search Result Cluster Labeling

Let the pages be represented by $p_1, p_2, \ldots . p_n$ where $p_i \in C$ and C is a cluster of search result containing n pages. To find all the related pages, first we initialize each page's list with pages that it connects.

$$l(p_i) = \cup (p_j), \forall 1 \le j \le n \;\&\; j \ne 1 \tag{1}$$

Using the heuristic, in each list, we search for the page, which connects to the largest number of pages in the cluster. To find such page, we look for the page which has the largest list. Let's denote such page as h.

$$h(p_i) = p_j : \arg\max \left\| l(p_j) \right\| \forall 1 \le j \le n \;\&\; j \ne i \tag{2}$$

After finding h page for each p_i, each list is updated by adding the pages of their corresponding h's list. This

process is performed until there is no change in the lists. The lists are then merged. This resultant list contains all the connected pages. These pages will contribute in defining the label of the cluster.

The page title is a good source of information about the page contents. A page title usually contains few words. To infer label from set of page titles, we use the famous Apriori algorithm for frequent itemset mining by Rakesh et al [16]. We treat each title as transaction and title words as items. Apriori algorithm is an iterative search method which uses k-itemsets to find k+1-itemsets. It also states that all subsets of a frequent itemset must be frequent. In our method, we set that value of k to 2 as labels having 2 words is sufficient to define a cluster.

Let's T be a set of all titles. We create a set of words W from the titles. All the stop words are removed from the titles. Let's each word of all the titles be considered as item w_i where $w \in W$, and all the titles as transactions t_i and $t \in T$. All the stop words are removed from the titles and are tokenized. We create an index of transactions and items as:

Table 1. Titles as Transaction and Title terms as Items

Transaction	Items
t_1	W_i, $0<i<=\|W\|$
...	...
t_n	W_i, $0<i<=\|W\|$

We specify minimum support count as 2. We consider the words which has minimum frequency of 2. First we create a set of frequent 1-itemset which satisfies the minimum support count. Then the set of frequent 2-itemsets is found by joining the set of 1-itemset with itself. All the itemsets in this set must satisfy the minimum support count. In this set, the most frequent 2-itemset is selected to be used in labeling the cluster.

The performance of our method depends on our heuristic search method. Our heuristic search method reduces complexity by avoiding non-promising nodes or pages at each iteration.

4. Dataset and Result

We select the famous "Jaguar" dataset for our experimentation. Since we are not concerned with the clustering process, we manually categories the first 100 hits of the Google search engine's result on the query "jaguar". We found 21 categories, out of which car, animal and sport categories are the largest.

Table2. Jaguar Dataset

Category	No. of pages	Category	No. of Pages
Car	55	Hotel	2
Animal	17	Photo Gallery	1
Sport	4	Movie	1
Super Computer	1	Touring	1
Music	1	Timing Systems	1
Music Band	2	Eyewear	1
Computer Game	1	Financial Firm	1
Emulator	1	Mining	1
Magzine	1	Scientific Prog. Package	1
Telecommunication Corp	1	Under Water Vehicle	1
Resin Models	1		

We treat each category as a cluster. In the car category, we found 15 pages that are hyperlinked using our heuristic search. In animal category, we found 3 and in sport, we found 2 pages having hyperlinks between them.

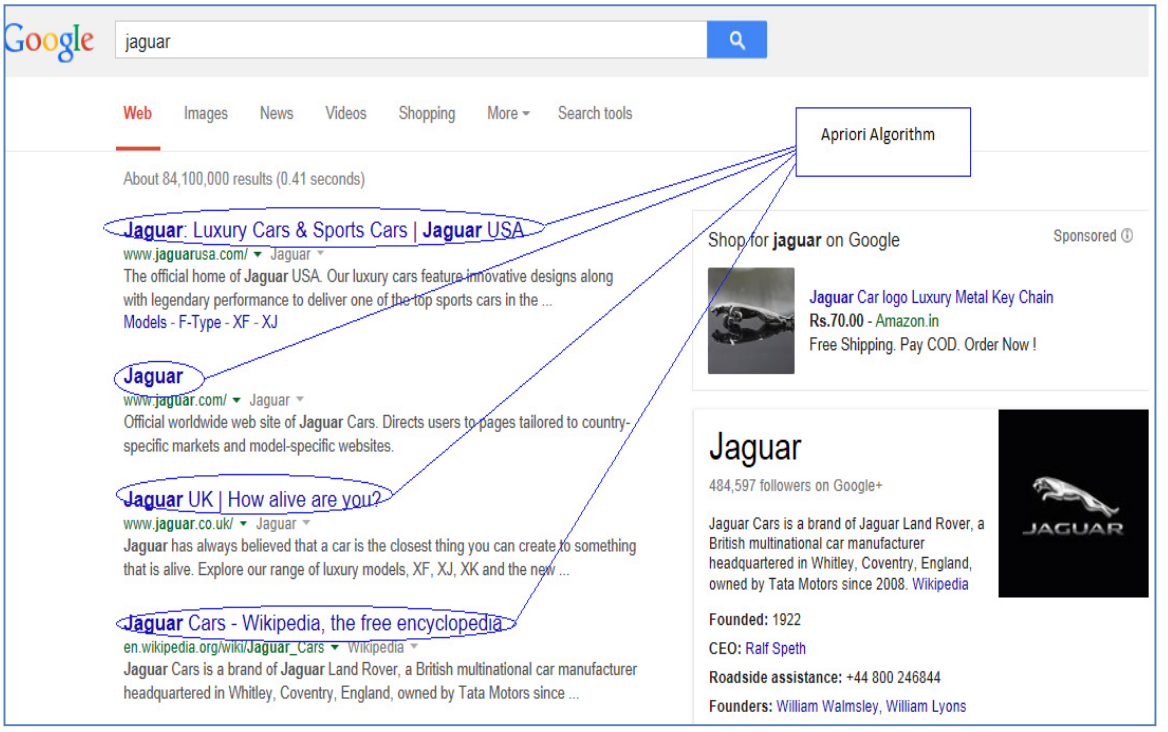

Fig. 1. Google's result on "Jaguar" query

The Google display of "jaguar" search result in Fig.1. is slightly different from our jaguar dataset as Google's result changes time to time depending on the frequently searched pages.

Table 3: Car Documents' Titles

Title #	Title
T1	Jaguar: Luxury Cars & Sports Cars | Jaguar USA
T2	Market Selector
T3	Jaguar UK | Jaguar
T4	Jaguar Cars
T5	Jaguar South Africa | Jaguar
T6	Jaguar Land Rover Careers; Excellence In Motion
T7	JAGUAR
T8	Jaguar Azerbaijan
T9	Jaguar Romania | Jaguar
T10	Jaguar Heritage
T11	New and Used Jaguar Dealer | Vancouver, British Columbia | Jaguar Vancouver
T12	Jaguar, latest Jaguar car news, reviews, pictures and videos - Telegraph
T13	Jaguar | Auto Express
T14	Jaguar Profile Instagram
T15	Jaguar Facebook

Table 4: Animal Documents' Titles

Title #	Title

T1	Jaguars, Jaguar Pictures, Jaguar Facts - National Geographic
T2	Jaguar Facts - Big Cat Rescue
T3	Jaguar - Wikipedia, the free encyclopedia

Table 5: Sport Documents' Titles

Title #	Title
T1	Jacksonville Jaguars, Official Site of the Jacksonville Jaguars
T2	Jacksonville Jaguars Football Clubhouse - ESPN

After tokenization and stop word removal process, we apply Apriori algorithm with minimum support 2 for frequent 2-itemsets to infer the labels. For the car cluster, we obtain the label "Jaguar Cars", which rightly reflects it. All the pages in the sport cluster are about the Jacksonville Jaguar team. Our method correctly infers the label for this cluster. For the animal cluster, we get the label "jaguar facts".

Table 6: Labels

Cluster	Label
Car	Jaguar Cars
Animal	Jaguar Facts
Sports	Jacksonville Jaguars

5. Conclusion

Labeling of clusters is as important as clustering. The labels should be meaningful and must convey the idea about the contents of the clusters. If the label, representing the cluster, is not appropriate, then there is less chance of selection by users. In this paper, we proposed a method to label clusters on the basis hyperlinks shared by their members and titles. The advantage of our approach is that instead of considering all the members to contribute in labeling process, we take those members' titles which share hyperlinks. These linked pages form the theme of the cluster. The pages' titles are then searched for frequent words. It saves us lots of computation and produces appropriate cluster labels. This work is in its early stage and we would try to make it more mature by applying it on other search result clusters dataset. Also we would try to include text within "meta" tag of each page in labeling process.

References

1. Li X, Chen J, Zaiane O. Text Document Topical Recursive Clustering and Automatic Labeling of a Hierarchy of Document Clusters. Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science Volume 7819, 2013, p 197-208.
2. Anaya-Sánchez H, Pons-Porrata A, Berlanga-Llavori R. A New Document Clustering Algorithm for Topic Discovering and Labeling. Progress in Pattern Recognition, Image Analysis and Applications, Lecture Notes in Computer Science Volume 5197, 2008, p.161-168.
3. Alfred R, Fun TS, Tahir A, On CK, Anthony P. Concepts Labeling of Document Clusters Using a Hierarchical Agglomerative Clustering (HAC) Technique. The 8th International Conference on Knowledge Management in Organizations Springer Proceedings n Complexity 2014, p.263-272.
4. Hanumanthappa M, Prakash BR, Mamatha M. Improving the Efficiency of Document Clustering and Labeling Using Modified FPF Algorithm. Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011, Advances in Intelligent and Soft Computing Volume 131, 2012, p. 957-966.
5. Nourashrafeddin S, Milios E, Arnold D. Interactive text document clustering using feature labeling. Proceedings of the 2013 ACM symposium on Document Engineering, p.61-70.
6. Thirunarayan K, Immaneni T, Shaik MV. Selecting Labels for News Document Clusters, NLDB, Paris, France, 2007, p. 119-130.
7. Popescul A. Ungar LH. Automatic Labeling of Document Clusters. Unpublished manuscript, available at:http://citeseer.nj.nec.com/popescul00automatic.html.
8. Carmel D, Roitman H, Zwerdling N. Enhancing Cluster Labeling Using Wikipedia, SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA, p.139-146.
9. Dostal M, Nykl M, Jezek K. Cluster Labeling With Linked Data, Journal of Theoretical and Applied Information Technology, July 2013. Vol. 53 No.3.
10. Fukumoto F, Suzuki Y. Cluster Labeling based on Concepts in a Machine-Readable Dictionary, Proceedings of the 5th International Joint Conference on Natural Language Processing, 2011, p. 1371–1375.
11. Zhang C, Xu H. Clustering Description Extraction Based on Statistical Machine Learning, Second International Symposium on Intelligent Information Technology Application, 2008.
12. Chen K, Liu L. ClusterMap: Labeling Clusters in Large Datasets via Visualization, CIKM'04, November 8-13, 2004.
13. Turel A, Can F. A New Approach to Search Result Clustering and Labeling, Information Retrieval Technology Lecture Notes in Computer Science Volume 7097, 2011, p.283-292.
14. Geraci F, Pellegrini M, Maggini M, Sebastiani F. Cluster Generation and Cluster Labelling for Web Snippets: A Fast and Accurate Hierarchical Solution, String Processing and Information Retrieval Lecture Notes in Computer Science Volume 4209, 2006, p.25-36.
15. Fernandes dos Santos F, Oliveira de Carvalho V, Rezend SO. Selecting Candidate Labels for Hierarchical Document Clusters Using Association Rules, Advances in Soft Computing Lecture Notes in Computer Science Volume 6438, 2010, p.163-176.
16. Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules, Proceedings of the 20th VLDB Conference Santiago, Chile, 1994.