ENHANCING THE LABELLING TECHNIQUE OF SUFFIX TREE CLUSTERING ALGORITHM

R.Mahalakshmi[1] and V.Lakshmi Praba[2]

[1]Research Scholar, Manonmaniam Sundaranar University, Tirunelveli,Tamilnadu
[2]Assistant Professor,Government arts College,Sivaganga,Madurai,India

ABSTRACT

Clustering the results of a search helps the user to overview the information returned. In this paper, we look upon the clustering task as cataloguing the search results. By catalogue we mean a structured label list that can help the user to realize the labels and search results. 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 method is introduced. More emphasis was given to /produce comprehensible and accurate cluster labels in addition to the discovery of document clusters. We also present a new metric that employs to assess the success of cluster labelling. We adopt a comparative evaluation strategy to derive the relative performance of the proposed method with respect to the two prominent search result clustering methods: Suffix Tree Clustering and Lingo.

we perform the experiments using the publicly available Datasets Ambient and ODP-239

KEYWORDS

Information retrieval, clustering, Search results clustering, Suffix Tree Clustering ,cluster labeling

1. INTRODUCTION

The overwhelming amount of textual documents available nowadays highlights the need for information organization and discovery. Effectively organizing documents into a hierarchy of topics and subtopics makes it easier for users to browse the documents. [1]

Search engines would retrieve all the documents based on the given key word and rank them according to the priority and display the documents page wise. Lets take top 20 ranked documents ,except the first 3 or 4 rest of the documents are not satisfying the user. More than 50 % of the documents are not relevant to the query since the retrieval is based on keyword and not based on the semantic similarity.

The ideal solution is , forcing the user to input a larger number of highly accurate keywords, trim down the number of results and yields somewhat improved ranking precision. This solution is, unfortunately, not practical for the normal user since most users tend to input not more than 3

DOI : 10.5121/ijdkp.2014.4104

keywords[3] . To solve this problem, Cluster the search result approach has been proposed. These methods have two main advantages. One is to make it easier for the user who has a clear search target to locate the desired document because the user can easily select the most appropriate cluster. The other is to assist the user, who would like to browse using just a few keywords or who has no clear search target, in understanding the outline of the search result through the labels of the clusters. The other important benefit is that the user often finds interesting information apart from what he/she wants to retrieve. Realizing these benefits, however, requires not only correct clustering but also labels clear enough to explain the clusters[5]. A lot of research is being intended towards clustering ,not using conventional clustering methods but regarding the clustering task as a task of selecting important key terms or phrases [5][6][7][8][10].To extract the terms as label candidates from documents firstly we consider that proper nouns are vital for characterizing documents. Second is a new label selecting criterion that can select the labels from candidates .The main objective of this paper is to provide more precise cluster label generation, superior group content discovery and incremental processing.

This paper deals with the enhancement of generalized Suffix tree based clustering approach. In general the most repeated phrase in the document tags is considered as cluster name. The general labelling procedure of the Suffix Tree algorithm is enhanced to improve the cluster label quality. This paper aims at organizing web search results into clusters facilitating quick browsing options to the browser providing excellent interface momentous labels to clusters .Suffix tree clustering produces comparatively more accurate and informative grouped results. The paper is organized as follows, Section1 deals with Introduction and section II briefs the related work ,section III describes the proposed methodology section IV details the experiments ,section V discusses the results and analysis ,finally section VI gives the conclusion of the paper.

2. RELATED WORK

Clustering is a common unsupervised learning technique used to discover group structure in a set of data. While there exist many algorithms for clustering, clustering is difficult because, unlike in supervised learning, there are no class labels for the data and, thus, no obvious criteria to guide the search. Another important problem in clustering is the determination of the number of clusters, which clearly impacts and is influenced by the feature selection issue.

Scatter/Gather [11] is one of the first system that dealt with cluster labeling, in addition to the cluster’s important terms, the titles of the document close to the centriod are also considered. **Filippo** et al [3] demonstrated that labels extracted from titles provide better description that those extracted from page’s content. There is a lot of research on linguistic-based summarization techniques for multiple documents which are also related to the labeling task. Radev et al.[10] [16]. However, multi-document summaries are usually too long to be utilized as short comprehensive labels. Several labeling approaches attempt to enrich terms by exploiting external resources for labeling, for example, the WordNet lexical database [4] was used to extract root meanings of important terms and to determine semantic relationships among these terms. **David Carmel** [10] utilized Wikipedia to represent the meaning of a text fragment as a weighted vector of Wikipedia concepts. Wikipedia has recently become one of the major knowledge resource for many information retrieval tasks, including text categorization and clustering [15, 16, 17], computing semantic relatedness between concepts [18, 9], and predicting document topics[10] [19]. Toda and Kataoka [21]also used named entities extracted from the text for labeling.

However, in many cases, a labeling approach that is solely based on the cluster content may have difficulties in providing discriminative labels. Several labeling solutions look for alternative resources.

3. SEARCH RESULT CLUSTERING ALGORITHM - STC

Zamir and Etzioni [4] introduced the suffix tree clustering algorithm (STC), which runs in $O(n)$ without computing $O(n2)$ similarity values. In detail, STC is made up of three steps

Step 1. A suffix tree for all suffixes of each document in $D = \{d1, \ldots, dn\}$ is constructed, and each suffix is associated with the set of documents wherein it is contained. In other words, using the notation given above, for each edge e (each of which represents a certain suffix) the set $S(e)$ is computed. The sets $S(e)$ with $|S(e)| \geq 2$ are called "base clusters" and identify the documents di with $i \in S(e)$.

Step 2. Each base cluster is assigned a score f, which is a function of $|S(e)|$ and the length of the suffix that is represented by e. In [Zamir and Etzioni 1998] the authors propose f as the product of $|S(e)|$ and the length of the suffix that is represented by e.

Step 3. The k base clusters $S1, \ldots, Sk$ that score best under f are selected. A similarity graph in which the base clusters form the node set is generated, and an edge between two nodes Si and Sj is added if the Jaccard coefficient of Si and Sj is larger than 0.5, say, when $|Si \cap Sj| \; |Si \cup Sj| > 0.5$. The connected components of this graph form the final clusters.

Step 4: Each base cluster is associated with a suffix, which can serve as a label for this cluster. This method solves two basic problems in topic identification for document clusters [Stein and Meyer zu Eißen 2004b]:word order preservation and topic length determination

STC [9] has proven to work well on document snippets that are returned by search engines , but its properties have been analyzed by many researchers . As pointed out above STC is a heuristic algorithm which is highly efficient, and has got few drawbacks.

4. IMPROVED STC ALGORITHM

After the execution of Step 3 in the normal STC ,that is before assigning base cluster suffixes to the clusters ,all the cluster label phrases are reviewed by the following criterions

1. If we have two cluster labels as synonyms or with the simple difference like singular and Plural merge the two into a single cluster ,discard the replicated documents and unique documents could be retained by selecting a unique label from the existing key phrases.

2. If any verb is given as a label for the cluster, relabeling it by adding a meaningful noun from the Key phrases or if not found, prefix the query phrase in to it.

3. On line databases are used to find the synonyms of the key phrases. Two online data collections are increasingly used in all kinds of ways in IR. One is Wikipedia. The other is WordNet. here we used Wordnet. Wordnet is used to find the parts of option which will list out

few Synonyms, then by applying term weight and term ranking we can select the most descriptive phrase/label .

4. If a cluster label is part of the another cluster label then their relevance is checked. If more than 50% of the documents are replicated/overlapped in both the clusters, the rest are checked against other cluster doc if all are overlapped with other cluster then remove the cluster, otherwise add them in to other topics. Hence we effectively lessen shadowing in the generated clustering.

5. EXPERIMENTS

ODP239, Ambient and PubMed datasets are used for our experiments. We run the two SRC's namely STC and LINGO with 10 different query key words and analyzed the same query with the enhanced STC.The general observations are represented by graphs. Table 1 shows the basic parameters and the attributes settings of the Suffix Tree algorithm ,Table 2 lists out the sample resultant cluster labels produced by LINGO and STC. Table 3 shows the revised list of improved cluster labels with no of documents. Number of Clusters are reduced from 16 to 12

Table 1 parameters setting of STC and Lingo algorithms

Dataset	ODP239	Ambient
Query Keyword	Women's health, jaguar Education,Musiums, Agriculture,environment,dictionary,entertainment	Montecarlo,Butterfly, camel,
No of Doc	100	71
No of Clusters	20	20
Algorithms	STC,LINGO	STC,LINGO
Max Phrases Per Label	4	4
Optimal Label Length	3	3
Base Cluster Merge Threshold	60%	60%
Word Boost	60-80%	60-80%
Max Cluster Phrase Overlap	60%	60%
Single Term Boost	80%	80%

Table 2 List of query terms, Datasets,Ambiguous lables and improved labels generated by STC, Lingo and enhanced STC

S.No	Query Term	Ambiguous labels in Lingo	Ambiguous Labels in STC	Revised Labels in Enhanced STC	Dataset
1	Cooking	Lemon,onion,includes	Lemon,uses,recipe, recipie	Soups,recipies	ODP239
2	Children	Kids,Child,designs,Available,sets	Baby,Babies,Offers,offeringKids,Children	Babies,Kids	

3	Entertainment	DVD,DVDS,specializing,region,selecion, merchandise,sales	selection,large selection of CDs, CDs,DVDs,CDs& DVDS	Selection,CDs,DVDs	
4	Women's health	Forum,answer,topic etc	Women,woman,female,topic,center,condition,treatment..	Health centers, diagnosis and Treatment uterus,prolapse,bladder	
5	Beagle	adopting, search	Linux,Desktopsearch,Club,national clubs,pages	Open source,National beagle clubs,	Ambient
6	Scorpian		Scorpian,Scorpians,Stung,Scorpian stings,Known	Scorpians,Scorpian stings	
7	Computer	increase,presence,selective,changes, improvement,efficient	Available,use,used, method,methods,significant,study	Compter uses,methods, significant performance	PubMed
8	Operating System	Score,organic,increased,water,biological,Parameters, surgery,fields	performance, analysis,clinical,used, study	performance analysis,clinical systems	
9	Nobel prize	changed,number, highly, reprogramming,	reprogramming,cell,new,conjugated,shared	Nobel prize shared	
10	Java	concentration,predicting, outcome,etc	indonesia.indonesian,high, compared,used	indonesia,outcome and analysis	

The following are the observations

1. Few Label names are found to be duplicated then the cluster is removed and the unique documents are merged with other cluster example "Treatment " is removed and the unique documents are merged with the " Diagnosis &Treatment"

2. Verbs are given as label name for few clusters, they are further prefixed by a query phrase and relabeled.(ex Label name CENTER is renamed as HEALTH CENTER to give more clarity to the user.) Wordnet online dictionary is used.

3. Some label names are synonyms, Ex label name FEMALE and WOMEN are the same ,the overlapping documents are removed and the remaining documents are merged with the WOMEN cluster..

4. label names are improved by increasing the wordboost by 100 % for example there was a lable called TOPIC and it was labeled by UTERUS,PROLAPSE,BLADDER. After increasing the wordboost to 100% ,which gives more precision about the group to the user.

5. Document overlapping is reduced by 12 % overall.

6. RESULTS AND ANALYSIS

6.1 Evaluation of Cluster Labels

Although clustering has been studied for several decades, the fundamental problem of a valid evaluation has not yet been solved. Evaluating the quality of clustering results is still a challenge in recent research. The sound evaluation of clustering results in particular on real data is inherently difficult. In the literature, new clustering algorithms and their results are often externally evaluated with respect to an existing class labeling**.[14].**The cluster labels could be evaluated based on the following parameters Comprehensibility Descriptiveness, Discriminative power Uniqueness, Non-redundancy.

1. Comprehensibility (f1):

A reader should have a *clear imagination* of the contents of a cluster. It can be formally defined as the following.$\forall c \in C \forall p \in lc: P \in L(G)^{p} > 1$ where lc is the cluster label of cluster c, p a phrase of lc, and L(G) determines a formal language identifying noun phrases.

$$\mathrm{f1(p) = NP(p) \cdot Penalty(p)},$$

1.0

$$\mathrm{NP(p)} = \begin{cases} 1 \quad , & if\ P \in L(G) \\ 0 \quad , & Otherwise \end{cases}$$

1.0

$$Penalty(p) = \begin{cases} exp\frac{-(p)-(popt)^{2})}{2.d^{2}}, & if\ p = 1 \\ 0.5 & otherwise \end{cases}$$

2.0 Descriptiveness (f2):

Every document of a cluster should contain the associated cluster label

$\forall c \epsilon C \exists p \in lc \forall p^{'} \in Pc: dfc(p') \ll dfc(p)$ where Pc is the set of phrases in the cluster c.

$$f2(c,p) = 1 - \frac{1}{Pc/Lc} \cdot \textstyle\sum_{p' \epsilon Pc} dfc(p^{'})/dfc(p) \qquad 2.0$$

3.0 Discriminative Power(f3)A cluster label should *only* be present in documents of its own cluster, could be formally defined as

$$\forall_{ci \neq cj}^{ci, cj} \in C \exists P \in lc: \frac{dfci(p)}{(ci)} \ll \frac{dfcj(p)}{(cj)}$$

criterion $f3(cj,p) = 1 - \frac{1}{k-1} \sum_{\substack{ci \in C \\ ci \neq cj}} \frac{cj.dfci(p)}{ci.dfcj(p)}$ 3.0

4.0 Uniqueness (f4):Cluster labels should be unique. , formally defined as

$$\forall \binom{cicj}{ci \neq cj} \in C: lci \cap lcj = 0$$

$$f4(cj,p) = 1 - \frac{1}{k-1} \sum_{\substack{ci \in C \\ ci \neq cj}} \frac{P \cap lci}{P \cup lcj} \qquad 4.0$$

5.0 Non-redundancy (f5)Cluster labels should not be synonymous, formally,
$\forall c \in C \forall p, p^{'} \in lc: p \textit{ and pi are not synonymous}$

$$f5(c,p) = 1 - \frac{1}{lc-1} \sum_{\substack{p \in lc \\ p^{'} \neq p}} Syn(p,p^{'})\ Syn : pxp - \{0,1\} \qquad 5.0$$

6.0 Relevance of a phrase with respect to a cluster:All constraints can be combined into a single criterion:

$$rel(c,p) = \sum_{i=1}^{F} wi.fi(c,p) \qquad 6.0$$

where wi is a weighting factor and F = {f|1 . . . 5}, namely,f1 Comprehensibility,f2 Descriptiveness,f3 Discriminative Power,f4 Uniqueness,f5 Non-redundancy .Note, that the effect of every constraint on the quality of a phrase is so far unevaluated.

Table 3 Label evaluation

Algorithm	**f1- Comprehensibility**	**f2 Descriptiveness**	**f3 Discriminative Power**	**f4 Uniqueness**	**f5 Non-redundancy**
Suffix tree clustering	0.70	0.73	0.86	0.84	0.90
Lingo	0.80	0.69	0.89	0.89	0.93
Enhanced suffix tree clustering	0.78	0.80	0.91	0.93	0.98

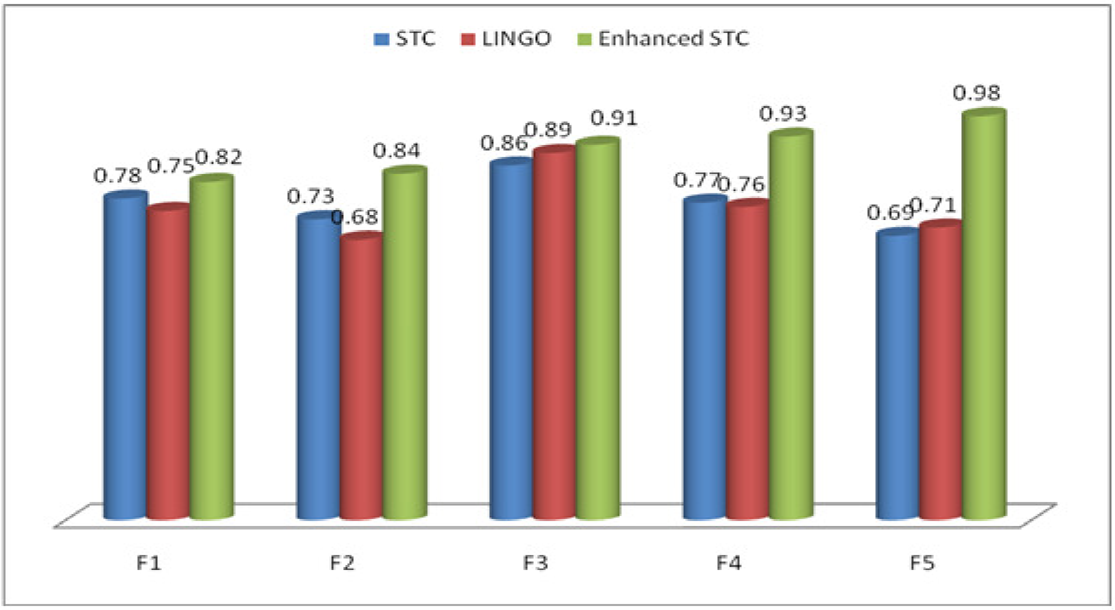

Fig 1. Comparative analysis on label evaluation of STC,LINGO and Enhanced STC

Table.4 Label relevance

S.No	Algorithm	LabelRelevance
1	STC	82.2
2	Lingo	80.2
3	Enhanced STC	89.6

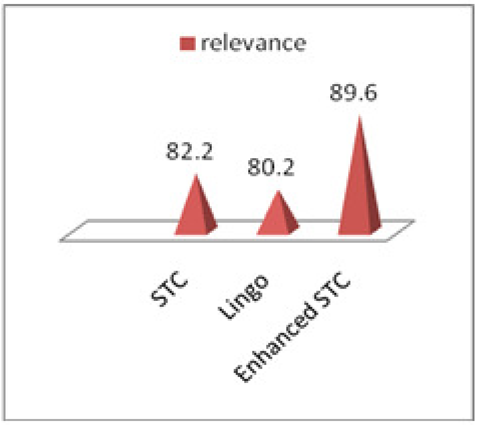

Fig 2. Comparitive analysis of Label relevance of STC,Lingo and Enhanced STC

7. CONCLUSION

Search results clustering is one of many methods that can be used to improve user experience while searching collections of text documents. We must accurately and concisely describe the contents of the cluster, so that the user can quickly decide if the cluster is interesting or not. This aspect of document clustering is sometimes neglected. cluster labelling is not less important than clustering. In fact a good cluster with a poor descriptive label is likely to be ignored by the user. As discussed that evaluating the quality of clustering results is still a challenge in recent research The cluster labels are evaluated based on the following parameters Comprehensibility Descriptiveness, Discriminative power Uniqueness, Non-redundancy. It was proved that the new proposed method of enhanced STC produced semantically meaningful, comprehensible and compact text labels to the document clusters. This paper addressed the effect of repeated acquisition of labels for search results clusters when the labelling was imperfect. We examined the improvement in label quality via avoiding repeated labelling, and focus especially on the improvement of descriptive and unique labels. The result shows that the Overlapping is reduced by 12 % and the labels relevancy is also improved by 8.4 percentage when compared to other two SRC's STC and LINGO. We are able to increase the label quality with respect to different parameters described in the evaluation criteria. If the key phrases are verbs that could be improved by finding appropriate noun in the key phrases or by prefixing the query term with the verb.In our work we didn't concentrate much on the label name as adjectives.

REFERENCES

[1] Text Document Topical Recursive Clustering and Automatic Labeling of a Hierarchy of Document Clusters Xiaoxiao Li1, Jiyang Chen, and Osmar Zaiane

[2] "Cluster Generation and Cluster Labelling for Web Snippets Filippo Geraci 1 ,2, Marco Pellegrini, Marco Maggini 2, and Fabrizio ebastiani

[3] A Search Result Clustering Method using InformativelyNamed Entities" Hiroyuki Toda NTT Cyber Solutions Laboratories, NTTCorporation WIDM'05, November 5, 2005, Bremen, Germany.Copyright 2005 ACM 1-59593-194-5/05/0011 ...$5.00.

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

[5] Zamir, O., Etzioni, O. and Grouper, A.: "Grouper: A Dynamic Clustering Interface to Web Search Results." roceedings of WWW8, pp.1361-1374, 1999.

[6] Zeng, H. J., He, Q. C., Chen, Z., Ma, W. Y. and Ma, J.: "Learning to Cluster Web Search Results." Proceedings of SIGIR'04, pp.210-217, 2004.

[7] Kummamuru, K., Lotlikar, R., Roy, S., Signal, K. and Krishnapuram, R.: "A hierarchical monothetic document clustering algorithm for summarization and browsing search results." Proceedings of WWW'04, pp.658-665, 2004.

[8] Ohta, M., Narita, H. and Ohno, S.: "Overlapping Clustering Method Using Local and Global Importance of Feature Terms at NTCIR-4 Web Task." Working Notes of NTCIR(NII-NACSIS Test Collection for IR Systems)-4 Vol.Supl. 1, pp.37-44, 2004.

[9] 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.uni-weimar.de) Martin Potthast (Paderborn University, Germany beebop@upb.de

[10] Enhancing Cluster Labeling Using Wikipedia David Carmel, Haggai Roitman, Naama Zwerdling IBM Research Lab Haifa 31905, Israel{carmel,haggai,naamaz}@il.ibm.com SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA. Copyright 2009 ACM 978-1-60558-483-6/09/07 ...$5.00.

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

[12] C. Clifton, R. Cooley, and J. Rennie. TopCat: Data Mining for Topic Identification in a Text Corpus. IEEE Trans. Knowl. Data Eng., 16(8):949–964, 2004.

[13] W. de Winter and M. de Rijke. Identifying Facets in Query-Biased Sets of Blog Posts. In Proceedings of ICWSM 2007, pages 251–254.

[14] On Using Class-Labels in Evaluation of Clusterings Ines Färber1, Stephan Günnemann1, Hans-Peter Kriegel2, Peer Kröger2, Ahornstrasse 55, 52056 Aachen, Germany, 2Ludwig-Maximilians-Universität München Oettingenstrasse 67, 80538 München, Germany.

[15] F. Geraci, M. Pellegrini, M. Maggini, and F. Sebastiani. Cluster Generation and Cluster Labelling for Web Snippets: A Fast and Accurate Hierarchical Solution. In Proceedings of SPIRE 2006, pages 25–36.

[16] S. Osi ´nski, J. Stefanowski, and D. Weiss. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In Proceedings of IIPWM 2004, pages 359–368.

[17] A. Popescul and L.H. Ungar. “Automatic labeling of document clusters”. http://www.cis.upenn.edu/~popescul/Publications/popescul00labeling.pdf, 2000.

[18] B. Stein and S. Meyer zu Eißen. Topic Identification: Framework and Application. In Proceedings of I-Know 2004, pages 353–360.

[19] H. Toda and R. Kataoka. A Clustering Method for News Articles Retrieval System. In Proceedings of WWW 2005, pages 988–989.

[20] D. Weiss. “Descriptive clustering as a method for exploring text collections”. Ph.D. dissertation. Pozna´n University of Technology, Poland, 2006.