Search Result Clustering Through Expectation Maximization Based Pruning of Terms

K. Hima Bindu and C. Raghavendra Rao

Abstract Search Results Clustering (SRC) is a well-known approach to address the lexical ambiguity issue that all search engines suffer from. This paper develops an Expectation Maximization (EM)-based adaptive term pruning method for enhancing search result analysis. Knowledge preserving capabilities of this approach are demonstrated on the AMBIENT dataset using Snowball clustering method.

Keywords Information retrieval · Clustering · FPtree · Expectation maximization

1 Introduction

Due to enormous growth of information available on the Web, search engine users are swamped with millions of search results, especially in case of broad and ambiguous queries. Ambiguity arises from the low number of query words [7]. Users typically give short queries with an average query length of three words per query [12]. Many search engines use diversification techniques to avoid duplicate results on the first pages of the results. This approach enables quick retrieval of one relevant result per subtopic, but may not facilitate retrieval of more results of user interest [13].

Search Results Clustering (SRC) is a solution to address the lexical ambiguity issue [2]. This approach is especially useful in case of polysemous queries. Search Results Clustering partitions the results obtained in response to a query into a set of labeled clusters. Each cluster corresponds to a subtopic of the query. Therefore, the user need not accurately predict the words used in the documents that best satisfy his

K. Hima Bindu (✉) · C. Raghavendra Rao
Department of Computer and Information Sciences, University of Hyderabad, Hyderabad, Andhra Pradesh, India
e-mail: himagopal@gmail.com

C. Raghavendra Rao
e-mail: crrcs@uohyd.ernet.in

B. V. Babu et al. (eds.), *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28–30, 2012*, Advances in Intelligent Systems and Computing 236, DOI: 10.1007/978-81-322-1602-5_134,

or her information needs. Instead, the user can start with a generic or broad query and quickly navigate to relevant subtopic. Further, user can grasp the semantic structure in the search results or refine the queries by using the cluster labels.

Search Results Clustering has specific challenges in contrast to traditional clustering methods. These include quick response time (as this is an online activity), ephemeral clustering (clustering happens when the user issues a query), meaningful cluster labels (must be human readable), and no fixed number of clusters.

A Frequent Itemset-based document clustering technique [4] can generate meaningful cluster labels and does not require the number of clusters as an input parameter. However, identification of frequent itemsets is a complex and time-consuming task and also involves subjectivity. Hence, we used Expectation Maximization (EM) [3] to prune the irrelevant terms to make our approach meet search result analysis challenges, in particular the SRC requirements. Snowball clustering proposed by [6] develops clusters starting with highest score (fitness) to least score. Hence, we have chosen to build a clustering method based on these approaches.

The outline of this paper is as follows. Section 2 briefly discusses the search results acquisition procedure and the preprocessing methods. The objective term pruning approach based on EM is presented in Sect. 3. Section 4 briefs the clustering and labeling algorithm. Section 5 shows our experimental results and the comparison with other popular data centric algorithms used for SRC. We conclude the paper in Sect. 6.

2 Search Results Acquisition and Preprocessing

The search results of a search engine are acquired by using the search engine's API by sending HTTP requests. All major search engines provide APIs with restrictions on the number of queries per day as a free service and as a paid service without such limitations. By sending a RESTful (Representational State Transfer) request to the public search engine APIs, the results are available in either JSON or XML format. Usually, the first 100 results are considered.

The title of each search result along with one or two lines summary of the web page (called as snippet) forms a search result document. These search results are usually preprocessed by tokenization, stemming (Porter Stemmer) and stopword removal. Stop words are frequently occurring words, which do not carry semantics. We used the stop word list of 571 words from SMART[1] system. Then, the words are stemmed by Porter's suffix stripping algorithm[2]. By using Vector space model, each result is represented as a TF-IDF vector [8]. The terms of a result's title and snippet after these steps form the features.

[1] http://www.lextek.com/manuals/onix/stopwords2.html

[2] http://tartarus.org/~martin/PorterStemmer/

3 Term Pruning with Expectation Maximization

Expectation Maximization is frequently used for data clustering in Machine learning, as it can handle latent variables in the data. Expectation Maximization method is developed for characterizing the population characteristics, which is a mixture of finite fundamental factors. Each factor will have its own uncertainty model characterized by associated probability distribution. The characteristics of the population will obey an uncertainty model characterized by mixed distribution, which is generalization of fundamental factors. This mixture model will have set of parameters, namely mixing proportions, and the parameters of fundamental factor distributions. If one assumes that the population is a mixture of k fundamental factors, and each factor obeys normal distribution, when the characteristic under study is one-dimensional, then the mixture model will have parameter's mixing proportions $\pi_1,\ \pi_2,\ \ldots \pi_k; \left(\pi_i > 0 \text{ and } \sum_{i=1}^{k} \pi_i = 1\right)$ and $(\pi_i,\ \sigma_i)$ as the mean and standard deviation of ith fundamental factor. Estimation process of these parameters is addressed by EM [3] based on the sample.

The terms occurring in each preprocessed search result constitute the population for the analysis. However, some of the terms are relevant and some are not relevant for building the knowledge. Thus, the distribution of the TF-IDF values of the terms can be viewed as a mixture of relevant and irrelevant terms (i.e., as a mixture of two distributions). The simpler way of representing it is by using Gaussian distribution, $f\ (x) = \pi_1 f_1\ (x|\mu_1, \sigma_1) + \pi_2 f_2\ (x|\mu_2, \sigma_2)$, where f_1 is the probability distribution function of TF-IDF values of relevant terms with mean μ_1 and standard deviation σ_1, $f_2$is the probability distribution function of TF-IDF values of non-relevant terms with mean μ_2 and standard deviation σ_2; π_1 and π_2 are the mixing proportions. These parameter estimates must have the property that $\mu_1 > \mu_2$. Based on the parameters obtained for the mixture model, the threshold on TF-IDF values is $\frac{\pi_1\mu_1+\pi_2\mu_2}{\pi_1+\pi_2}$, from [14]. Any term with TF-IDF value more than this threshold is classified as relevant term, otherwise irrelevant term.

4 The Clustering and Labeling Algorithm

Search Results Clustering algorithms must be faster to have minimum latency between query submission and cluster presentation. Our algorithm performs clustering and labeling on the fly by processing only the search results without using external knowledge, hence, it is a lightweight approach. It labels the clusters automatically by using only the text concepts (frequent termsets) available in the search results.

The frequent termsets are identified by FPGrowth [5]**,** by considering the relevant terms retained after the EM based term pruning, snowball clustering method [6] has been applied to arrive at the clusters along with the labels. The size of the relevant

terms set due to EM makes the frequent itemset analysis manageable, though it is a tedious task.

5 Results and Analysis

An illustration of EM-based term pruning is provided here, by considering the first ten results of the query "Beagle," from the AMBIENT[3]. dataset.

5.1 Terms Extraction

The first 10 results after tokenization (using white space as delimiter) resulted in 189 terms, are shown in Sect. 5.1.1. With the preprocessing by stemming and stop-word removal as discussed in the Sect. 3, 118 terms are retained, these are shown in Sect. 5.1.2. The TF-IDF vectors are constructed for these terms. When EM algorithm is run, the threshold on TF-IDF values is obtained as 1.17, based on the method described in Sect. 4. It is observed that 28 terms' TF-IDF satisfy this threshold. These are treated as relevant terms, which are given in Sect. 5.1.3.

5.1.1 Terms Obtained After Tokenization

(EC), (software), -, –, ..., 2, 3, 5, :, A, ANSI/ISO, American, Apple, Architecture, BEAGLE, Beagle, Beagle, Beagle-type, Beagle, Beagles, Breed, Britannica, British, C++, C++, CenterÂ®, Club, Computation, Consortium, Debian, Desktop, Dog, Download, Earth, Encyclopaedia, English, Evolutionary, Files, GNOME, God's, Google, Group, Guide, Head, Hound, Information, Information, Kennel, Linux, Main, Mars, Open, Owner's, PSSRI, Package, Page, Pillinger, Professor, Profile:, Search, Size, Spotlight, Standard, Storage, The, There's, University-based, University, Unix, W3, Wikipedia, Windows, a, all, also, among, an, and, any, architectures, are, article, as, at, available, be, beagle, beagle, blog-post, breed, built, but, by, can, centuries, child, coat, code, coded, color, come, compliant, critters, cutest, data, dog, dog, dogs, easy-care, encyclopedia, enhance, entirely, existed, experience, exploration, fairly, for, framework, free, get, good, gotten, green, group, has, have, head, heavy, hound, idea, in, indexing, industrial, is, its, led, long, medium, member, modern, not, of, on, or, other, over, overwhelmed, pack, page, partners, people, personal, pet, point, post, power, product, puppies, puppy, reference, researchers, search, short, should, similar, simplicity, single, sized, skull, sleek, small, solidly, sporty, standard, system, the, their, this, to, together, tool, website, where, which, will, with, you, your, â€¢.

[3] AMBIguous ENTries : http://credo.fub.it/ambient.

5.1.2 Terms After Pre Processing by Stemming and Stop-Word Removal

american, ansiiso, appl, architectur, articl, avail, beagl, beagletyp, blogpost, breed, britannica, british, built, center, centuri, child, club, coat, code, color, come, compliant, comput, consortium, critter, cutest, data, debian, desktop, dog, download, earth, easycar, ec, encyclopaedia, encyclopedia, english, enhanc, entir, evolutionari, exist, experi, explor, fairli, file, framework, free, gnome, god, good, googl, gotten, green, group, guid, head, heavi, hound, idea, index, industri, inform, kennel, led, linux, long, main, mar, medium, member, modern, open, over, overwhelm, owner, pack, packag, page, partner, peopl, person, pet, pilling, point, post, power, product, professor, profil, pssri, puppi, refer, research, search, short, similar, simplic, singl, size, skull, sleek, small, softwar, solidli, sporti, spotlight, standard, storag, system, togeth, tool, univers, universitybas, unix, w3, websit, wikipedia, window.

5.1.3 Terms After EM-Based Term Pruning

american, architectur, britannica, british, club, code, debian, desktop, download, encyclopaedia, encyclopedia, explor, free, guid, inform, kennel, led, main, mar, open, owner, page, profil, puppi, search, softwar, w3, wikipedia.

5.2 Analysis

For the query "Beagle", by considering the first 100 search results, the number of terms after stemming and stop-word removal is 840. With term pruning, the number of terms is 230. The tokenized termset size is not reported as it not the interest of researchers.

The average terms set size for all the 44 queries of AMBIENT dataset, resulted without term pruning as 865.18 (with standard deviation 74.06), and with term pruning as 274.22 (with standard deviation 72.22). The percentage of reduction in terms set size is 68.30.

The associated FP Trees of pre and post term pruning have been constructed and the corresponding estimates of the size (obtained by multiplying the maximum depth and maximum width) are derived and reported in Fig. 1. For the 44 datasets of AMBIENT, the Mean (Standard Deviation) of FPTree sizes before and after term pruning are 2486.48 (SD = 286.12) and 962.82 (SD = 243.29), respectively. The percentage of reduction in FP tree sizes is 61.28.

The "degree of knowledge representation" (Kappa) developed in the Rough Set theory [10] which quantifies the knowledge representation in considered features is employed, treating the terms as features. It is observed that pre pruning produced mean Kappa as 1 (SD = 0), indicating that the pre pruned termset possess hundred percent knowledge about the ground truth. Whereas post pruning term set has mean Kappa 0.95 (SD = 0.04) indicates that, the knowledge representation by post pruned

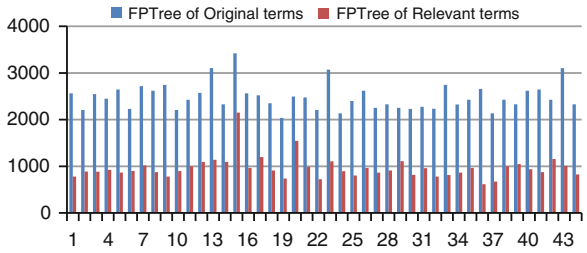

Fig. 1 FP Tree sizes of the AMBIENT dataset

term set will possess the knowledge about the ground truth from 0.8 to 1 (with mean 0.95). The relevant terms set may lead at the most 20 % loss in knowledge representation, 68.02 % gain by reduction in term set size and with 61.28 gain by FPTrees size reduction. This suggest that the relevant term set obtained by EM-based term set pruning can be employed in search result analysis in Web Mining.

The search result clustering method discussed in Sect. 4 has been considered for demonstrating the effectiveness of the recommended method.The distribution of the no of clusters with and without pruning is given in Fig. 2. The term pruning approach resulted in more specific clusters, so the number of clusters has increased.

RandIndex (RI) [11] is used as the evaluation measure, to compare our approach against the ground truth. The Rand Index values of each query when the clustering method is run with and without term pruning, are reported in Fig. 3. With term pruning, every query's RI value improved (statistically significant from our experiments).

When the Snowball clustering algorithm is run on the AMBIENT datasets, without term pruning, the average Rand Index has a low value of 58.83. When the term pruning is performed, RI value raised to 61.72. Comparison of our approach against some of the data centric SRC approaches, which do not use external resources is presented in Table 1. Lingo [9] uses Singular Value Decomposition, STC [15] uses Suffix trees and KeySRC [1] is based on Key Phrases.

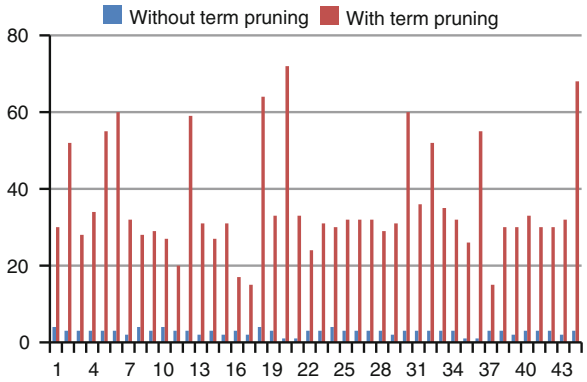

Fig. 2 Number of clusters with and without term pruning

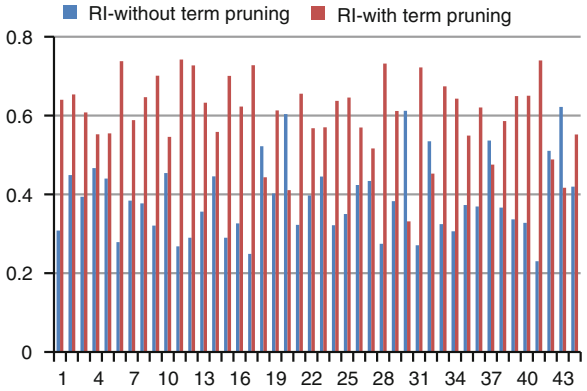

Fig. 3 RI values with and without term pruning

Table 1 Average rand index

Clustering method	Rand index
Lingo	62.75
STC	61.48
KeySRC	66.49
Snowball without term pruning	58.83
Snowball with EM based term pruning	61.72

6 Conclusion

The data centric SRC is effective due to the pruning of terms beyond well-defined stop words, through the EM algorithm. The search results are clustered and labeled with Snowball clustering method, without any human interaction or even external knowledge. As the threshold computed by EM algorithm is adaptive, the recommended method is adaptive machine learning technique. EM based term pruning has shown two benefits—overcome the curse of dimensionality and less runtime memory (FP Tree sizes are low).

References

1. Bernardini, A., Carpineto, C., D'Amico, M.: Full-subtopic retrieval with keyphrase-based search results clustering. In: Proceedings of Web Intelligence 2009, IEEE Computer Society, pp. 206–213 Milan (2009)
2. Carpineto, C., Osinski, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM. Comput. Surv. (CSUR) **41**(3) (2009). Article No. 17. ISSN:0360–0300
3. Dempster, A.P., Laird, N.M., Rubin, D.R.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc.: Ser. B (Methodol.) **39**(1), 1–38 (1977)
4. Fung, B., Wang, K., Ester, M.: Hierarchical document clustering using frequent itemsets. In: Proceedings of SIAM International Conference on Data Mining, pp. 59–70 (2003)

5. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generations. In: Proceedings of ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA (2000)
6. Hima Bindu K., Raghavendra Rao C.: Association rule centric clustering of web search results. In: Proceedings of MIWAI'11, Vol. 7080/2011, pp. 159–168, Hyderabad (2011). doi:10.1007/978-3-642-25725-4_14
7. Kamvar, M., Baluja, S.: A large scale study of wireless search behavior: google mobile search. In: Proceedings of CHI'06, pp. 701–709, NewYork (2006)
8. Manning, C. D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University press, New York, USA (2008)
9. Osinski, S., Stefanowski, J., Weiss, D.: Lingo: Search results clustering algorithm based on singular value decomposition. In: Proceedings of the International Intelligent Information Processing and Web Mining Conference. Advances in Soft Computing, Springer, 359–368 (2004)
10. Pawlak, Z.: Rough Sets—Theoretical Aspects of Reasoning about Data. Kluwer, Boston (1991)
11. Rand, W. M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
12. Scaiella, U., Ferragina, P., Marino, A., Ciaramita, M.: Topical clustering of search results. In: Proceedings of WSDM 2012, pp. 223–232. ACM, New York (2012)
13. Taghavi, M., et al.: An analysis of web proxy logs with query distribution pattern approach for search engines. Comput. Stand. Interfaces. 162–170 (2011). doi:10.1016/j.csi.2011.07.001
14. Wajahat Ali, M. S., Raghavendra Rao, C., Bhagvati, C., Deekshatulu, B. L.: EMTrain++: EM based incremental training algorithm for high accuracy printed character recognition system. Int. J. Comput. Intell. Res. **5**(46), 365–371 (2010)
15. Zamir, O. And Etzioni, O.: Web document clustering: A feasibility demonstration. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, 46–54 (1998)