Term Importance Degree Impact on Search Result Clustering

Soheila Karbasi
Golestan University, 49138-15739 (155)
Gorgan, Iran

Mehdi Yaghoubi
Golestan University, 49138-15739 (155)
Gorgan, Iran

ABSTRACT
As wellactual clustering algorithms have to deal with explosive growth of documents of various sizes and terms of various frequencies, an appropriate term-weighting scheme has a crucial impact on the overall performance of such systems.

Term-weighting is one of the critical process for document retrieval and ranking in most search result clustering systems.In this paper we introduce a new technique forclustering algorithms that solve the problem of indexing the terms of big datasets and their characteristicswhich exist in most of current clustering approaches.The paper focus on term frequency normalization step ofclustering algorithms. Anew factor has been applied *tobasic term-weighting schemes for using in clustering process. T*he evaluated results confirm the impact of this factor to increase the performance of clusteringtechniques.The experiments were carried out on the standard algorithms and ODP-239 datasets which validated by statistical tests.

General Terms
Information system, Data Mining, Search result clustering

Keywords
Weighted clustering,Term importance degree, Term frequency normalization

1. INTRODUCTION
Todays, continuously increasing the number of text documents,intranets and digital libraries on the web leads to the use of more efficient search engines and retrieval methods. The results of search engines are shown as a snippet consisting of a title, URL and small text excerpt from the selected source website. When a user transmits a query to a clustering based search engine, the engine employs its indexing structure and retrievesrelevant results for the query. The results are deliveredto the search resultclustering system andconsequently, retrieval clustering mechanism starts to operate.Search result clustering system filters the text data and extracts important features. It clusters andlabels the input snippets according to its algorithm and offers labeled groupsof results. Finally, clustered search results are presented from a web interface tobe reviewed by users [1, 2].
Aperfect search result clustering output is composed of thematic grouping relatedto the given query with meaningful and representative labels of the groups. Usersread each label at a glance and naturally estimate the coverage of snippets insidethat cluster. They decide whether results in each cluster is in accordance with theinformation need without looking inside the cluster. If they explore the clustercontents by clicking on the label, the snippets inside should satisfy the informationneed or at least increase the knowledge of users about the query.Clusters containdocuments about a subtopic of the query and each cluster is labeled to give information about the subtopic which guide users and decrease search time during their search process. In fact, finding the underlying subtopics of search results returned for a queryis a hard task. Even, manually clustering and labeling is a complexand time consuming work, so automatic solution of this problem is still open forimprovement [3].

Many clustering approaches have been proposed and studied in the domain of textual information retrieval so that term-weighting process should provide an indicator of importance to discriminate the terms for labelingthe clusters [4].

In this paper,we present a new search result clustering method based on term weighting amelioration which is one of the most important steps in all of the clustering algorithms. We aim to cluster search results accurately by redefining the representation of documents and the weights that assigned to their indexed terms.To offer new scheme, we utilize some clustering evaluation metrics used in literature. The implementationsof these evaluation metrics, namely, weighted F-measure and normalized mutual information are also adapted by Carrot2 API [5].

The main goal of new scheme is to emphasise the importance of distribution of document size in datasets. It was realised that a particular term with a high frequency is not necessarily in a long document which means the term frequency will be penalized by classic methods of normalization methods [6].

In order to evaluate the performance of presented scheme for clustering and labeling, it hasbeen tested with ODP-239 test collection and compared with baseline results. The rest of the paper is organized as follows: Section 2 reviews the popular clustering models and focuses on performance measures used for clustering and labelingtasks. Section 3, proposes the experimental settings and proposed method. Experiments and results are described in Section 4 and Section 5 contains the discussions and conclusionswith possible future pointers.

2. SEARCH RESULT CUSTERING
Several models are proposed in the literature for search result clustering. The methodology used in these models are to extract the relationships among websites and to construct the final clusters through feeding theresults.

In spite of the recent progresses in usualtechniques, the performance of text based clustering systems is largely dependent on term-weighting models. Typically, clustering algorithms use the Vector Space Model (VSM) [7] to encode documents. The VSM relates terms to documents, and since different terms have different importance in a given document, a term weight is associated with every term [8]. These term weights are often derived from the frequency of a term within a document or set of documents. Many term weighting schemes have been proposed [9, 10]. In addition, large-scale retrieval performance requires the use of appropriate term-weighting scheme since it dominates the computational demands of retrieval [11].

One of the most commonly used term-weighting schemes is tf-idf and its variants [8].One common characteristic of tf-idfweighting schemes is that they all require knowledge of the entire indexed terms of the collection. Hence, it is evidence that with increasing the number of documents of datasets, any applications that rely on the indexed terms will be affected. This problem will be emphasis in the case of web search results.

In this paper, we use a new term weighting scheme, based on BM25, which generates document representations based on term importance degree[12]. Afterward, calculated weights are used in the indexing phase of clustering techniques.

In fact, the previous experiments shows that all of the terms within a document are not good indicators for document content and it is better to consider more significant terms for document discrimination[13]. Besides, it is examined that when the size of dataset is small, the numberof unique terms continues to climb up as the number of documents increases. However, this growth is reduced sharply as the number of documents becomes verylarge. This observation indicates that if a dataset is sufficiently large (for example, documents on the web), we can expect to see very few new terms by adding more documents [6].Therefore, in search result clustering, it is important to determine more significant terms of documents and websites and they should be relied much more than the others in term-weighing schemes.

In the next section, our suggestedterm weightingschemeis introduced and its influence in STC, Lingo and K-Meansclustering algorithmsare presented.

3. EXPERIMENT ARCHITETURE

All of our clustering experiments are applied with three originalclustering algorithms (STC, Lingo and K-Means)used in Carrot2 API.

STC is a linear time clustering algorithm based on identifying common phrasesto all documents. A phrase is defined as an ordered sequence of one or more terms[14]. Its difference from other clustering algorithms is that STC considers a sentenceas a sequence of connected terms instead of common bag of terms usage. Clustering is performed using common phrases between documents usingsuffix tree data structure.

Lingo [15] is a well-known successor of STC which frequent phrases are extracted using suffix arrays instead of suffix trees.Next, the frequentphrases that best match certain latent topics present in the search results which are determined via singularvalue decompositionare selected and finally documents areallocated to such frequent phrases.

K-Means is one of the most common and popular algorithms published first by J. B. Macqueen in 1967. From the algorithm's name, it's required to specify a K number of desired clusters. Then, the algorithm randomly selects K snippets as initial seeds for search result clustering. Next, the algorithm assigns the rest of the snippets to the closest seeds and calculates the new cluster centroids by taking the average value for every dimension. The algorithm repeats the process of calculating new cluster centroids untilclusters' boundaries become stable. There are two major downsides of K-Means algorithm. It is non-overlapping algorithm where snippetscannot belong to more than one cluster. Also, it is sensitive to outliers [16].

Term weighting is one of the imperative step in all of the clustering algorithms which can be altered by associated algorithm. This difference is based on two main method (document-centred and cluster-centred approaches) of clustering algorithms.

In our work, we proposed that there isn't a significant meaning and relation between term frequency anddocument frequency and total number of document's terms. We measured the importance degree of termswhich is determined by ranking the terms based on term frequency within each document as a significant factor to apply in term-weighting scheme. This factor assesses the importance of terms not only by frequency, but also by frequency rank. Hence, we used functions 1and 2to calculate the weights of terms used in the term weighing step of clustering algorithms.

$$W = \frac{tf_{ij} \times log\left(\frac{N - df_i + 0.5}{df_i + 0.5}\right)}{(0.5 + 1.5 \times Avg\_Rank_j) + tf_{ij}} \quad (1)$$

$$Avg\_Rank_j = \frac{\sum_{i=1}^{|d_j|} tid_{ij}}{|d_j|} \quad (2)$$

Where:

tf_{ij}*:* frequency of term i in documentj

$|d_j|$: number of unique terms in document*j*

df_i*:* number of documents containing term i

N: total number of documents in dataset

tid_{ij}: importance degree of term i in document j

Avg_Rank_j*:* average of terms importance degree of document*j*

It is observed that number of terms per document has an important role in term frequency normalization. Beside, distribution of documents according to their lengths in larger datasets will be more widespread. Proposed parameter called Avg_Rank, normalize term frequency and represents the average of importance degree of document terms.

The main advantage of Avg_Rank is that, it is unique in each one of the documents and independent from the others. The intuition behind this setting is that, using term importance degree has a positive impact in term-weighting scheme and it will be useful to establish a specific term frequency normalization parameter for each document. The small value of Avg_Rank in a document means that the majority of document terms are important and this parameter must increase the weight of these terms. Contrarily, if the value of this parameter is high, it means that the document has a lot of terms and therefore its score will be reduced for clustering.

In order to assess the performance of the proposed algorithm,we performedour experiments in one of the publicly available datasets specific to SRC task:ODP-239 dataset [17]. The ODP-239 dataset consists of 239 queries, each with 100 snippets andabout 10 subtopics. Each search result consists of a URL, title and a very short text. The dataset is derived from Open Directory Project (ODP) [18].Table 1 shows statistical characteristic of dataset.

Table 1. ODP239 characteristic

# Query	239
# Snippet	25580
Avg. # snippet per query	107
Avg. # cluster per query	9.5
Avg. # snippet per cluster	11
Avg. # term per snippet	19

4. EXPERIMENTS AND ANALYSIS

This section provides the evaluation results of our approach to clustering tasks in test dataset described in Section 3. The F-measure and NMI(Normalized mutual information) metrics obtained with original used term-weighting function in Carrot2(log tf-idf) and function 1are presented in Tables2 and 3.

As it is seen,obtainedF-measureswith all of the algorithms have been increased.As the same as F-measure, in the term of NMI metric, the results show that using new scheme of discrimination the termsachieve great improvement in purity of clusters.

It can be concluded that the proposed weighing scheme based on term importance degree produce better discrimination between snippets which can be used in any search result clustering algorithms. The most advantage of this method is that, reducing the number of indexed terms will be more effective in the case that the size and number of documents increase dynamically.

Table 2. F-measuresby different clustering algorithms in the case of using original and new term-weighing schemes

Algorithm	Initial F-measure	New F-measure
STC	0.510	**0.590**
Lingo	0.430	**0.484**
K-Means	0.458	**0.512**

Table 3. NMI by different clustering algorithms in the case of using original and new term-weighing schemes

Algorithm	Initial NMI	New NMI
STC	0.416	**0.490**
Lingo	0.480	**0.503**
K-Means	0.403	**0.449**

5. CONCLUSION

The main goal of this work was based on decreasing the number of indexed terms used in all of text clustering algorithms. The aim was to use an auxiliary parameter for better normalization frequency of terms which plays the principal role in retrieval process especially in large and heterogeneous datasets.

The intuition behind the presented approach was that, all of the terms within a document are not good indicators for document topic and it is appropriate to consider more significant terms for discriminationthe documents from each other. Besides, this will be very suitable for clustering and labeling the documents via decreasing the number of comparison and total required time. Hence, it is important to determine significant indexed terms and they should be relied much more than the other terms in term-weighing schemes.Consequently, this selection can increase clustering efficiency and decrease total time of clustering task which is the aim of any search result clustering engine.The experiments on test dataset have revealed that the proposed scheme haspositive performance onF-measure and NMI metrics, and it is significantly faster than classic methods.

6. REFERENCES

[1] H. J. Zeng, Q. C. He, Z. Chen, W. Y. Ma, and J. Ma. Learning to cluster websearch results. In Proceedings of the Special Interest Group on InformationRetrieval (SIGIR) Conference on Research and Development in InformationRetrieval, pages 210–217, Sheffield, United Kingdom, 2004. ACM Press.

[2] D. Zhang and Y. Dong. Semantic, hierarchical, online clustering of websearch results. In J. X. Yu, X. Lin, H. Lu, and Y. Zhang, editors, Asia-PacificWeb Conference, volume 3007 of Lecture Notes in Computer Science, pages69–78. Springer, 2004.

[3]C. D. Manning, P. Raghavan, and H. Schutze. Introduction to InformationRetrieval. Cambridge University Press, 2008.

[4]Salton, G. & McGill, M.J., Introduction to Modern Information Retrieval. McGraw-Hill, New York 1983.

[5]S. Osinski, J. Stefanowski, and D. Weiss. Lingo: Search results clusteringalgorithm based on singular value decomposition. In Intelligent InformationSystems, pages 359–368, 2004.

[6]Joel W. R. et al., TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams, *ICMLA, pages 258-263. IEEE Computer Society,* (*2006*).

[7] Salton., G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley, 1989.

[8] Salton, G. & Buckley, C., Term-Weighting Approaches in Automatic Text Retrieval, Information Processing & Management, 24(5), pp. 513-523, 1988.

[9] Salton, G., Syntactic approaches to automatic book indexing. In Proc of the annual meeting on Association for Computational Linguistics (ACL) (1988), pages 204-210, Department of Computer Science, Cornell University, Ithaca, New York, 1988.

[10] Anh, V. & Moffat, A., Simplified similarity scoring using term ranks, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, August 15-19, 2005, Salvador, Brazil.

[11] Baeza-Yates, R., &Ribeiro-Neto, B., Modern information retrieval. Harlow, England: Addison - Wesley Longman Ltd, 1999.

[12]Robertson, S., Walker, S., M. M. Beaulieu, Gatford, M. & A. Payne, Okapi at trec-4. In NIST Special Publication 500-236: The Fourth Text Retrieval Conference (TREC-4), pages 73 - 96, 1995.

[13]Karbasi, S. &Yaghoubi, M., International Journal of Computer Applications, Volume 38, January 2012.

[14] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In Proceedings of the Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieval, 1998. [15]Osinski, S., Weiss, D., 2005. A Concept-Driven Algorithm for Clustering Search Results. IEEE Intelligent Systems, 20 (3), 48–54.

[16] A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[17] C. Carpineto and G. Romano. Odp-239 dataset. http://credo.fub.it/odp239/, 2009. Accessed on August, 19, 2011.

[18] Open directory project. http://www.dmoz.org/. Accessed on August, 19,2011.