Keyword Clustering for Automatic Categorization

Qinpei Zhao Mohammad Rezaei Hao Chen
Pasi Fränti
School of Computing, University of Eastern Finland
qinpei.zhao@uef.fi

Abstract

Processing short texts is becoming a trend in information retrieval. Since the text has rarely external information, it is more challenging than document. In this paper, keyword clustering is studied for automatic categorization. To obtain semantic similarity of the keywords, a broad-coverage lexical resources WordNet is employed. We introduce a semantic hierarchical clustering. For automatic keyword categorization, a validity index for determining the number of clusters is proposed. The minimum value of the index indicates the potentially appropriate categorization. We show the result in experiments, which indicates the index is effective.

1. Introduction

With the development on Internet, web-based and *information retrieval* (IR) applications, such as search engines, social networks, multi-media sharing, customer reviews are exploded. Short texts such as search query, comments, photo description and tags are the modern means in the applications. Although text classification and clustering are well studied, the techniques are not successful in dealing with short texts. The short text are typically lack of context information, free form and highly unstructured. Thus processing short texts is challenging. To enrich the short texts representations, external resources such as WordNet [1], Wikipedia [2] and Google search results [1, 2, 4, 6, 7, 12] get involved.

Search engine queries are mostly short texts. The average length of them is about 2.3 terms and 30% have a single term [9]. A method grouping search results based on different meanings of the query is proposed [4] for efficiently identifying relevant results. To get a better semantic similarity, search engine results are employed [12, 1, 2]. For each pair of short texts, they do statistics on the results returned by a search engine (e.g., Google) in order to get the similarity score.

New inspired clustering algorithms have been proposed to deal with short text clustering. A framework of comments-driven clustering for organizing web resources is explored in [5]. The clustering approach is studied over the popular video sharing site *YouTube*. A probabilistic framework, which includes a knowledgebase (*Probase*) and certain inferencing techniques on top of the knowledgebase is proposed in [11]. The framework is to enable machines to perform human-like conceptualization. Experiments are conducted on conceptualizing textual terms and clustering short pieces of text such as *Twitter* messages. Also novel uses of validity indexes have been presented [3, 8]. An evaluation of different internal clustering validity indexes is presented to determine the possible correlation between the measures and F-measure [8].

Hierarchical clustering commonly employed in text clustering, is a method of cluster analysis which seeks to build a hierarchy of clusters. It provides *dendrogram* as clustering results. Non-hierarchical procedures usually require the user to specify the number of clusters before any clustering and hierarchical methods routinely produce a series of solutions ranging from one cluster to n clusters (assume n objects in the data set). Numerous methods for determining the number of clusters have been proposed for numerical data [10]. However, there is little research on validity index for keyword clustering.

In this paper, a new validity index, which determines the number of clusters for semantic hierarchical clustering is proposed. The method is applied for automatic categorization. Our focus is on strings with single word based on which processing on strings with multiple words are applicable. Since single words lack of content for statistical conclusion, we employ WordNet to get semantic similarity directly. The main contribution of this paper is to introduce a new validity index in

[1] http://wordnet.princeton.edu
[2] http://www.wikipedia.org

keyword clustering.

2. Semantic Hierarchical Clustering

Given a list of keywords $S = \{s_1, s_2, ..., s_n\}$, keyword categorization is to cluster them into groups, where the keywords in each group are semantically similar. The clusters are defined as $C = \{c_1, c_2, ..., c_k\}$. Hierarchical clustering can provide categorization with one to n clusters, i.e., $k = n$.

A semantic hierarchical clustering requires a measure of semantic similarity between data. The similarity measure can be obtained from external resources, such as Wikipedia and WordNet thesaurus. We use WordNet in this paper. Information-content based similarity measures such as Resnik, Lin and Jiang & Conrath are considered. Take an example of Jiang & Conrath in distance metric, which is defined as:

$$\begin{aligned} P(s) &= \frac{\sum_{w \in Set(s,s')} count(w)}{N} \\ IC(s) &= -\log P(s) \\ LCS(s,s') &= \max_{c \in Set(s,s')} IC(c) \\ JC(s,s') &= (IC(s) + IC(s')) - 2LCS(s,s') \end{aligned} \tag{1}$$

where $Set(s, s')$ is a set of words subsumed by s and s'. $P(s)$ is the probability that a random word (w) in the corpus is an instance of s. N is the number of words in the corpus. $LCS(s, s')$ (Least Common subsumer) is the lowest common ancestor node of s and s' in the hierarchy of WordNet.

An example of semantic hierarchical clustering result by Jiang & Conrath is shown in Fig. 1.

3. Automatic Categorization

In most real life clustering situations, an applied researcher is faced with the dilemma of selecting the number of clusters in the final result. Thus, a validity index for determining the number of clusters is necessary. The index is based on the dendrogram with cluster size one to n obtained from hierarchical clustering (see Fig. 1). It is used to decide at which level of the hierarchy the categorization is the best.

For getting a proper number of clusters, a fixed range of $[k_{min}, k_{max}]$ is usually pre-defined. It is meaningless to set $k_{min} = 1$ because uniform test (deficiency of randomness) is enough. Also clustering algorithm has no effect on one cluster. Thus, usually one sets $k_{min} = 2$ and $k_{max} \leq n$.

The index is defined based on the *Compactness* and *Separation* of clusters, which are defined as:

$$\begin{aligned} C(k) &= \max_t \{\max_{i,j} JC(s_i, s_j)_{s_i \neq s_j \in c_t}\} + I_1/n \\ S(k) &= \frac{\sum_{t=1}^{k} \sum_{s>t}^{k} \min_{i,j} JC(s_i, s_j)_{s_i \in c_t, s_j \in c_s}}{k(k-1)/2} \end{aligned} \tag{2}$$

Where, $C(k)$ represents compactness within clusters and $S(k)$ is separation between clusters. In $C(k)$, s_i and s_j are the ith and jth string in tth cluster c_t and I_1 is the number of clusters with one item. Similarly, s_i and s_j are the ith and jth string in tth cluster c_t and sth cluster c_s respectively, k is the number of clusters at that hierarchical level.

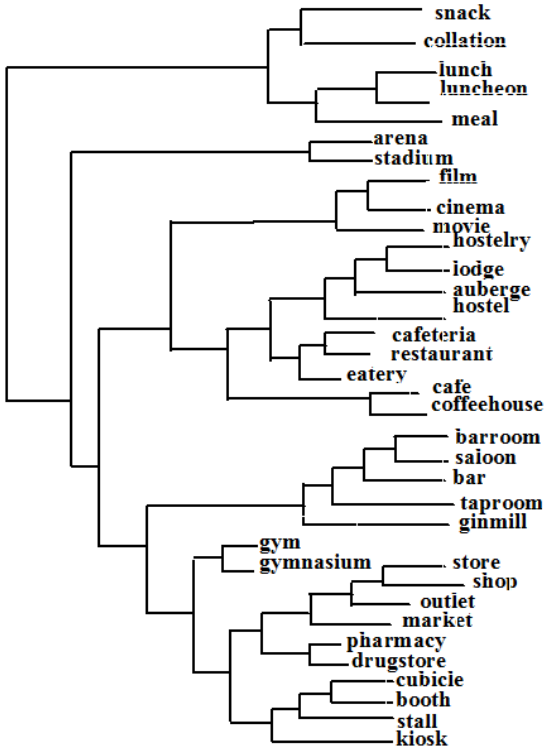

Figure 1. An example of dendrogram from semantic hierarchical clustering on data mopsi.

There exists a special case that cluster size is one, which means there is only one item in a cluster. For clustering, the special case is not preferred. And it is not possible to calculate the pairwise distance with only one item. Thus, we constraint the $C(k)$ by adding I_1/n. The categorization is assumed to be items within a cluster are as similar as possible and items between clusters

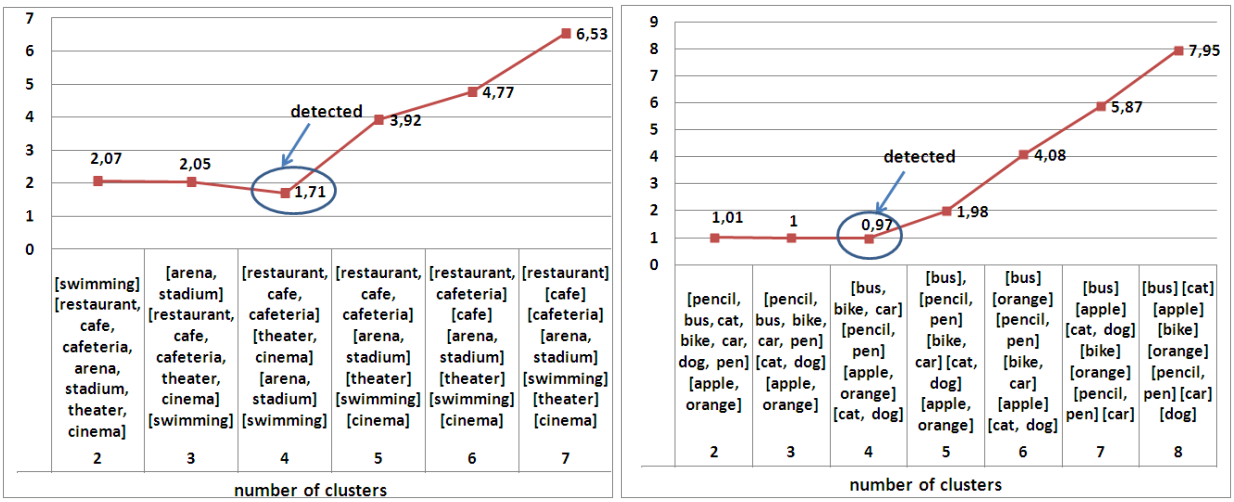

Figure 2. The stopping criterion on artificial data. Four is the minimum value for both cases.

are as different as possible. For the clustering result with k clusters, the validity index is defined as:

$$SC(k) = \frac{C(k)}{S(k)} \tag{3}$$

The index is calculated for each k among $[k_{min}, k_{max}]$. The k with minimum value in the range is selected as the best fitting number of clusters.

4. Experiments

The experiment is conducted on artificial data (see Fig. 2) and data mopsi obtained from *MOPSI*[3] project. The MOPSI project implements different location-based services and applications such as mobile search engines, photos, user tracking and route recording. The project has its applications integrated both on the web and mobile phones with the aim to integrate user location as a search option. The words in data mopsi (see Fig. 1), which contains 36 nouns, are picked up from services, search query keywords and photo descriptions. Since there are many unstructured words in Finnish language, we select a small sample and translate them into English by Google Translate API. We use Java to access the semantic similarity measures from WordNet 3.0. The user interface is programmed in JSP (Java Server Pages).

The validity index on artificial data is shown in Fig. 2. The x-axis is the number of clusters k and y-axis is the value of $SC(k)$. The categorizations are also displayed. The numbers of clusters detected by the stopping criterion are both four, where the categorization is reasonable from human judgment.

[3] http://cs.joensuu.fi/mopsi

For the real data mopsi, a ground truth categorization is obtained by 20 people. There are two persons who divide the data into 11 clusters, five persons into 10 clusters and 13 persons into 8 clusters. The dendrogram by the semantic hierarchical clustering is shown in Fig. 1. The number of clusters detected by the proposed validity index is nine (see Fig. 3), where the values of seven, eight and ten clusters are quite close. The categorization of nine groups is shown in Fig. 4. The maximum distances ($JC(s, s')$) within clusters and the minimum distances between clusters are displayed.

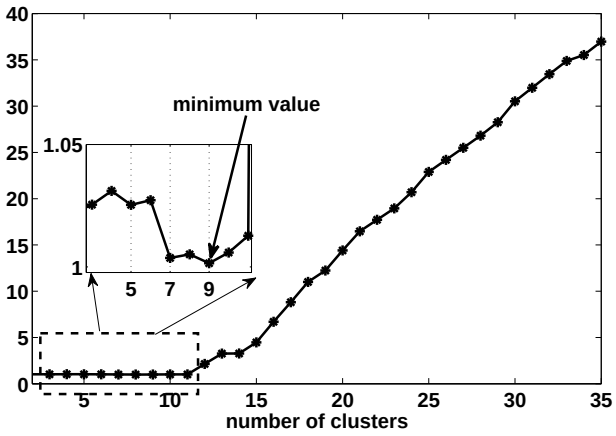

Figure 3. The validity index on data mopsi. Nine is the minimum value.

Even a number of clusters can be determined by an algorithm based on certain criterion, human judgment often differs from each other on the categorizations and the number of clusters. However, the proposed criterion

can suggest a potentially appropriate categorization.

The study is simply performed on nouns. It can be extended to verbs also. For strings with multiple words, the processing can be based on processing for strings with single words. However, it is more complicated to analyze strings with multiple words by WordNet.

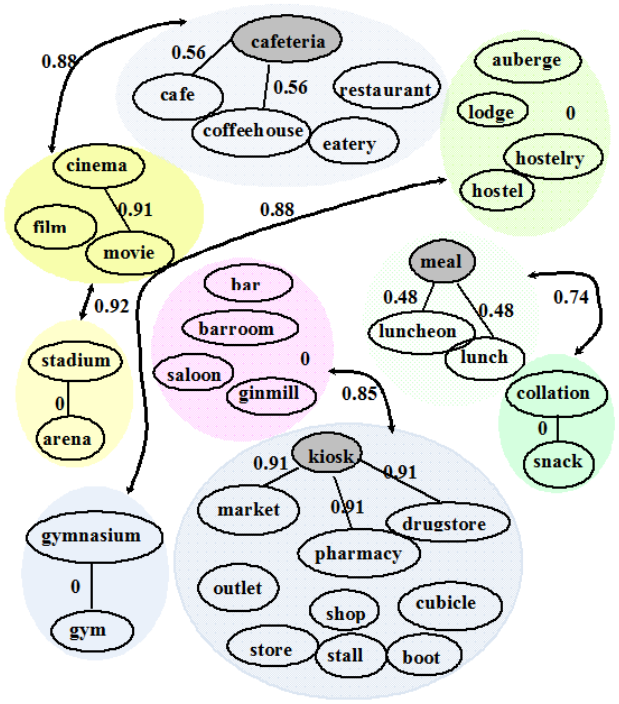

Figure 4. Categorization of nine groups on data mopsi with the minimum distances within clusters and maximum distances between clusters.

The semantic similarity obtained from WordNet sometimes has difference with human's judgment, which leads the undesired clustering result. For example, the similarity between words *lion* and *tomcat* is 0, however, the similarity between *lion* and *cancer* is 0.05. The hierarchical clustering merges *lion* and *cancer* as a group firstly, which does not match with human's judgment. Therefore, automatic categorization on the undesired clustering result is not reliable.

5. Conclusion

We introduced a keyword clustering for automatic categorization. For getting a semantic similarity, we employed the similarity measures from WordNet. A validity index in semantic hierarchical clustering was proposed for automatic categorization. The index is based on the compactness and separation of clusters, where the minimum value indicates a good categorization. The experiment performed in a real project indicates the method is working. Finding a better way to calculate semantic similarity for strings with either single word or multiple words is our future work. It is also interesting to study on other clustering algorithms, such as spectral clustering on this problem.

References

[1] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. *Proc. WWW*, pages 757–766, 2007.

[2] R. Cilibrasi. The google similarity distance. *IEEE Trans. on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[3] M. Errecalde, D. Ingaramo, and P. Rosso. A new AntTree-based algorithm for clustering short-text corpora. *Journal of Computer Science and Technology*, 10(1):1–7, 2010.

[4] R. Hemayati, W. Meng, and C. Yu. Semantic-based grouping of search engine results using wordnet. *APWeb/WAIM'07*, pages 678–686, 2007.

[5] C. Hsu, J. Caverlee, and E. Khabiri. Hierarchical comments-based clustering. *Proc. of the 2011 ACM Symposium on Applied Computing*, pages 1130–1137, 2011.

[6] J. Hu, L. Fang, Y. Cao, H. Zeng, H. Li, Q. Yang, and Z. Chen. Enhancing text clustering by leveraging wikipedia semantics. *Proc. of the 31st annual intl. ACM SIGIR conf. on Research and development in information retrieval*, pages 179–186, 2008.

[7] X. Hu, N. Sun, C. Zhang, and T. Chua. Exploiting internal and external semantics for the clustering of short texts using world knowledge. *CIKM'09*, pages 919–928, 2009.

[8] D. Ingaramo, D. Pinto, P. Rosso, and M. Errecalde. Evaluation of internal validity measures in short-text corpora. *CICLing'08*, pages 555–567, 2008.

[9] B. Jansen, B. Spink, J. Bateman, and T. Saraceric. Real life information retrieval: a study of user queries of the web. *ACM SIGIR Forum*, 32(1):5–17, 1998.

[10] G. Milligan and M. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[11] Y. Song, H. Wang, Z. Wang, and H. Li. Short text conceptualization using a probabilistic knowledgebase. *TechReport:MSR-TR-2011-26*, 2011.

[12] W. Yih and C. Meek. Improving similarity measures for short segments of text. *Proc. AAAI*, 2:1489–1494, 2007.