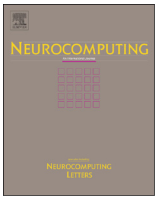
Neurocomputing 147 (2015) 472–484

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Automatic multi-way domain concept hierarchy construction from customer reviews

Ding Tu, Ling Chen*, Gencai Chen

College of Computer Science, Zhejiang University, Hangzhou 310027, PR China

ARTICLE INFO

Article history:
Received 16 January 2014
Received in revised form
22 April 2014
Accepted 7 June 2014
Communicated by Lijun Tang
Available online 27 June 2014

Keywords:
Taxonomy construction
WordNet
Hierarchical clustering
Taxonomy comparison

ABSTRACT

A concept hierarchy is important for many applications to manage and analyze text corpora. In the literature, most previous hierarchy construction works are under the assumption that the semantic relations in the concept hierarchy can be extracted from a text corpus, which is not fully satisfied for short and informal texts, e.g. tweets and customer reviews. And many works utilize hierarchical clustering methods to get the final concept hierarchy, in which the resulting binary-tree form concept hierarchy cannot fit the demand in many applications. In this paper, we propose a general process for building a concept hierarchy from customer reviews with an appropriate depth. The process can be divided into three steps. First, all highly ranked topic words are extracted as concept words using a topic model. And a word sense disambiguation task is performed to derive the possible semantics of the words. Then, the distances between these words are computed by combining their contexts and relations in the WordNet. Finally, all words are organized using a modified multi-way hierarchical clustering method. In addition, a new concept hierarchy evaluation model is presented. Our approach is compared to approaches using hierarchical clustering methods on the Amazon Customer Review data set, and the results show that our approach can get higher similarity scores with the reference concept hierarchy.

1. Introduction

As the information technology develops, the amount of documents grows rapidly. Some corpora are too big to be fully explored by users, e.g. the customer reviews about a particular product. Analyzing such a large amount of textual information without a certain extent of abstraction is a time-consuming work. An appropriate concept hierarchy can aid users to manage and analyze a huge corpus efficiently, e.g. the Amazon product Browse Tree Guide (BTG). However, the BTG is a hierarchy of all products, which means it cannot work well with the limited attributes and flat structure in a specific product domain. For example, there are about thousands of kinds of LCD TV products on Amazon and each has many customer reviews. For some popular products, there may be thousands of customer reviews. Since the manual maintenance of all the attributes of every product would cost a lot, automatic domain concept hierarchy methods are needed.

A concept hierarchy or a taxonomy can be defined as a specific form of an ontology, which is a formal, explicit specification of a shared conceptualization [6]. With a concept hierarchy, users can organize information into categories and concentrate on a particular aspect of the information. For example, a concept hierarchy can help users understand the main contents of the customer reviews and the relations between them efficiently.

The challenges of constructing an appropriate concept hierarchy from a corpus lay in three sides. The first challenge is to ensure the words in the concept hierarchy are the most typical and relative words for that corpus. After the words are determined, finding the semantic relations hidden in the text written by natural language is the second challenge. The last challenge is to organize the words into a hierarchy that reflects the semantic relations and is easy to be understood by users.

Extracting terms from a corpus is the first step to construct a domain concept hierarchy. The corresponding methods can be divided into three types. The first kind of methods employs statistical features to extract important words from the corpus. One feature used frequently is the term frequency inverse document frequency (TF-IDF). The method proposed by Sun et al. [27] considers this feature as one metric for concept extraction. The drawback is that pure statistical method may introduce many meaningless words. The second kind of method uses linguistic patterns to extract concept words, e.g. the method proposed by Cimiano et al. [5]. The limitation of pure linguistic methods is that the importance of a word cannot be measured. The last kind of

* Corresponding author. Tel.: +86 13606527774.
E-mail address: Lingchen@zju.edu.cn (L. Chen).

http://dx.doi.org/10.1016/j.neucom.2014.06.038
0925-2312/© 2014 Elsevier B.V. All rights reserved.

methods combines the advantages of the methods mentioned above; e.g. the method proposed by De-Knijff et al. [6] exploits six different filters to extract terms.

There have been many works proposed in the literature to construct a concept hierarchy and to extract semantic relationships from a set of concepts. Some works focus on using statistical methods to build a concept hierarchy, such as the methods presented in [13,17,5]. This kind of methods usually extracts statistical patterns to obtain semantic relations between words. Another kind of method uses predefined rules to obtain particular kinds of relations, e.g. the methods presented in [12,32]. These methods can reach higher accuracy but rely on the quality of the rules. Additionally, some works utilize external semantic dictionaries to extract semantic relations; e.g. Sun et al. [27] and Tu et al. [30] utilized WordNet [22] to help construct the concept hierarchy. The advantage of this kind of method is that it can work with a small corpus well.

These works mainly concern with the extraction of the semantic relations between the words, and the statistical features used to extract words are simple. Besides that, a large number of methods use hierarchical clustering and output hierarchies in a binary tree form. This means if the number of concept increases, the depth of the tree will increase rapidly, and the efficiency of the hierarchy will decrease in many cases. For example, the concept North America has four sub-concepts: USA, Canada, Greenland, and Mexico. It will be better if these four concepts are set under the concept North America as its children, which cannot be accomplished in a binary tree.

In this paper, we present a new approach based on concept context and WordNet to build a multi-way concept hierarchy from a document corpus. The proposed approach exploits a topic model to extract important concepts from a domain corpus, and applies a multi-way hierarchical agglomerative clustering algorithm to generate the final concept hierarchy.

Our contributions are as follows:

(1) Propose a new approach to construct a domain concept hierarchy from customer reviews.
(2) Present a distance metric that combines the semantic distances and the context distance, which adapts to the review data.
(3) Apply a modified multi-way agglomerative clustering algorithm to organize concept words extracted in the previous step.
(4) Propose a new metric to evaluate the similarity of two hierarchies with the same leaf nodes, and compare our approach with existing hierarchical clustering methods.

The rest of the paper is organized as follows: Section 2 introduces related work of the paper; Section 3 describes the process of our approach; the new evaluation model is presented in Section 4; Section 5 gives the experiments and the results are discussed; and Conclusions and future work are presented in Section 6.

2. Related work

2.1. Semantic relation extraction

There are three types of methods used in most academic articles to extract semantic relations. The first type uses the manmade semantic dictionaries, e.g. WordNet [23,27]. According to Rada et al. [25], the distance of the concepts on a semantic net has great correlation with human judgment of the semantic relatedness. The drawback of these types of methods is that they might be ineffective in an uncommon domain, because the words in such a domain may not be in the semantic dictionaries or the right sense of a word may not be in the semantic dictionaries. The second type of methods exploits the syntactic relations or the contexts of words in a corpus to obtain semantic relations, e.g. the methods presented in Cimiano et al. [5] and Kang et al. [15]. They employed formal concept analysis (FCA) and concept lattices to represent the syntactic dependencies acquired from a corpus and employed these to construct a concept hierarchy. The method proposed by Maher et al. [21] also uses FCA as their base to build a web services semantic lattice that enables the visual browsing of web services. This kind of methods relies on the quality and quantity of the corpus. The third type of methods utilizes rule-based extraction mechanisms. The Probase project [31] belongs to this kind and employs predefined rules to find “is–a” relations in a large corpus. Another example is the approach presented in [3]; it applies a weakly supervised rule induction algorithm to Wikipedia to extract instances of arbitrary relations. The disadvantage of rule-based extraction methods is that these methods may be not useful for some corpora which are not large and formal, and the quality and quantity of the rules are also important factors for them.

2.2. Hierarchical clustering

A hierarchical tree supports a multi-level view of a data set. Based on the forwarding direction of the hierarchy construction process, we can divide the hierarchical clustering methods into two kinds. The first kind of methods is hierarchical agglomerative clustering, which constructs the hierarchy from bottom to up. According to the scheme employed to measure the distance of two clusters, it can be further divided into three types: single-link, complete-link, and average-link [33]. Except the basic cluster distance metric, the DiSH [1] uses subspace clustering to group the data nodes, with density as the cluster distance metric. It is based on the idea that several subspace clusters of lower dimensionality may form a subspace cluster of higher dimensionality. Blundell et al. [4] proposed the Bayesian rose tree, which interprets the candidate trees as the mixtures over partitions of a data set. Another kind of method is hierarchical partitional clustering algorithms, which build the hierarchy from the opposite direction compared to the agglomerative methods. According to the criterion function used to optimize the entire clustering process, the partition-based methods can also be further divided, e.g. the criterion functions in Ding et al. [7] and Puzicha et al. [24].

2.3. Concept hierarchy similarity

To evaluate concept clustering methods, the similarity between the trees should be computed. This problem has been researched in many areas. Tree edit distance, a common method used in many researches, is first introduced in [29]. Goddard and Swart [10] propose a method to get distances between graphs through elementary edge operations, which is analogous to tree edit distance. Since the nodes in different concept hierarchies are not the same, concept hierarchy similarity is different from tree similarity to some extent. Therefore, other factors besides tree similarity, e.g. node similarity, should be considered while computing concept hierarchy similarity. There are some works on comparing the concept hierarchy similarity. Alexander Maedche and Steffen Staab [20] propose methods measuring ontology similarity from three levels: semiotic, syntactic, and pragmatic. Cimiano et al. [5] applied a core ontology model to define the similarity of two concept hierarchies in which they employ some ideas from Maedche. The main part of the evaluation model is the taxonomic overlap, which compares the taxonomy of two ontologies.