Knowledge-Based Systems 74 (2015) 1–13

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Adaptive Concept Resolution for document representation and its applications in text mining

Lidong Bing [a], Shan Jiang [b], Wai Lam [a], Yan Zhang [c,*], Shoaib Jameel [a]

[a] Key Laboratory of High Confidence Software Technologies, Ministry of Education (CUHK Sub-Lab), Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong
[b] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, United States
[c] Department of Machine Intelligence, Peking University, China

ARTICLE INFO

Article history:
Received 28 February 2014
Received in revised form 21 July 2014
Accepted 6 October 2014
Available online 1 November 2014

Keywords:
Adaptive Concept Resolution
Ontology
WordNet
Wikipedia
WordNet-Plus

ABSTRACT

It is well-known that synonymous and polysemous terms often bring in some noise when we calculate the similarity between documents. Existing ontology-based document representation methods are static so that the selected semantic concepts for representing a document have a fixed resolution. Therefore, they are not adaptable to the characteristics of document collection and the text mining problem in hand. We propose an Adaptive Concept Resolution (ACR) model to overcome this problem. ACR can learn a concept border from an ontology taking into the consideration of the characteristics of the particular document collection. Then, this border provides a tailor-made semantic concept representation for a document coming from the same domain. Another advantage of ACR is that it is applicable in both classification task where the groups are given in the training document set and clustering task where no group information is available. The experimental results show that ACR outperforms an existing static method in almost all cases. We also present a method to integrate Wikipedia entities into an expert-edited ontology, namely WordNet, to generate an enhanced ontology named WordNet-Plus, and its performance is also examined under the ACR model. Due to the high coverage, WordNet-Plus can outperform WordNet on data sets having more fresh documents in classification.

1. Introduction

Traditionally, the representation of text documents is usually based on the Bag of Words (BOW) approach, which represents the documents with features as weighted occurrence frequencies of individual words. This technique has several drawbacks. First, it breaks a phrase, say " air conditioner", into independent features. Second, it maps synonymous words into different features. Third, it merges a polysemous word's different meanings into a single feature. These drawbacks make the document similarity unable to be computed by BOW accurately. The methods that overcome these drawbacks can be categorized into two classes, namely, linear projection models (including LSA [7], PLSA [17], LDA [4], OPCA [32]), and S2Net [45], and ontology-based methods [19,35]. In this paper, we focus on the latter methodology.

Some expert-edited ontologies include WordNet [29], Cyc [27], Mesh [50], etc. Previous empirical results have shown some improvement in some applications utilizing ontologies [1,5,12,13,19,25,26,35,38,40,50,52]. Recently, the online collaborative encyclopedia Wikipedia[1] provides us another resource to assist the text mining tasks, and its potential has been shown in classification [14,47,48], clustering [20,21,30,36], semantic relatedness computing [44], and taxonomy induction [2,9,34,33]. However, the existing works have an obvious shortcoming: the strategies they adopted are static. For example, one strategy is to use each synset in the WordNet as one dimension in the representation vector of the documents. Therefore, the resolutions for representing the documents belonging to different collections are the same. Suppose we have two document collections, the first one has coarse granularity categories, such as sports and military, while the second one has finer granularity categories, such as football and basketball. In the first collection, football players and basketball players should be regarded as related, while in the second they should be unrelated. So an adaptive strategy is very likely able to outperform the static

* Corresponding author. Tel.: +86 1062755592.
E-mail addresses: ldbing@se.cuhk.edu.hk (L. Bing), sjiang18@illinois.edu (S. Jiang), wlam@se.cuhk.edu.hk (W. Lam), zhy@cis.pku.edu.cn (Y. Zhang), msjameel@se.cuhk.edu.hk (S. Jameel).

[1] http://en.wikipedia.org.

http://dx.doi.org/10.1016/j.knosys.2014.10.003
0950-7051/

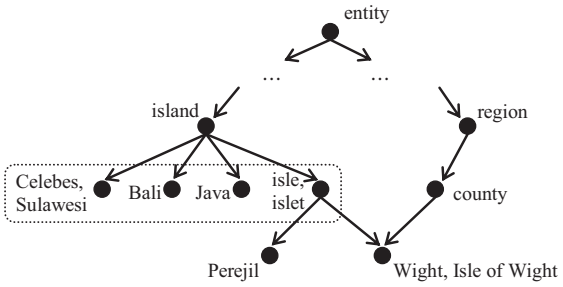

Fig. 1. A fragment of WordNet structure. Each node is a concept, whose synset contains the terms attached to the node.

one. Furthermore, in the existing works only one ontology is employed, either expert-edited one or online collaborative one. Hence they suffer from the former's limited coverage or the latter's noisy information.

In this paper, the proposed Adaptive Concept Resolution (ACR) model can learn a concept border from an ontology taking into the consideration of the characteristics of the particular document collection. Then, this border can provide a tailor-made semantic concept representation for a document coming from the same domain. The structure of an ontology is a hierarchical directed acyclic graph[2] (refer to the example in Fig. 1), and the border is a cross section in the graph. All the concepts located below the border will be merged into one of the concepts on the border. We design a gain value to measure whether a concept is a good candidate for the border. The gain value is calculated based on the characteristics of the given document collection. As a result, our model can generate different tailor-made borders for different collections adaptively. Another advantage of ACR is that it is applicable in both classification task where the groups are given in the training document set and clustering task where no group information is available. To do so, we only need to change the granularity, that is either cluster or individual document, for calculating the gain value. Therefore, ACR can be applied to both classification and clustering. The experimental results show that our model can outperform an existing static method in almost all cases.

Currently, there are more than 4 million English articles (i.e., entities) in Wikipedia, which makes it an extremely valuable linguistic repository. Wikipedia's ability of covering new terms is much better than expert-edited ontologies. Take the term "Bing" as an example, it may refer to a Web search engine from Microsoft, or a soft drink from UK, or others. But this term is not covered by WordNet. However, the abundant information is also a double-edged sword. Because Wikipedia is collaboratively edited by large number of users with different backgrounds and editing capabilities, it involves large amount of noise and its structure is very complicated. To leverage the advantages and eliminate the limitations, we propose a method to merge Wikipedia entities into the structure of an expert-edited ontology, i.e. WordNet, and construct an enriched ontology, called WordNet-Plus. Consider a Wikipedia entity, with the category information of the entity as clues. We can get a set of WordNet concepts which are the potential higher-level semantic meanings of the entity. Then, the similarity between the entity and each candidate concept is calculated to find the most suitable higher-level semantic meaning for the entity. Finally, we attach this Wikipedia entity under the found WordNet concept. Thus, WordNet-Plus keeps WordNet's good structure, meanwhile it encapsulates large amount of information from Wikipedia. Therefore, WordNet-Plus inherits the advantages of both WordNet and Wikipedia. In our experiment, 611,161 Wikipedia entities are integrated into WordNet. For example, a small island "Bacan" in Indonesia is successfully attached under the WordNet concept "island", the search engine "Bing" is attached under "search engine" and "website".

For comparing the performance of WordNet-Plus with the expert-edited ontology, both of them are applied to our proposed ACR model to generate two different representations for the same document. These two representations are applied to two different text mining tasks, namely, classification, and clustering. The results show that the performance of WordNet-Plus in text mining is competitive under ACR model compared with WordNet. In the Web page classification task, WordNet-Plus can outperform WordNet significantly because of its high coverage on new terms. In the clustering experiment, WordNet-Plus performs as good as WordNet on three data sets.

The presented work in this paper substantially extends our previous short paper [3] in several aspects. First, we elaborate the technique details of the proposed ACR model, which cannot be fully given in the short paper [3]. Second, we present a method to integrate Wikipedia entities into an expert-edited ontology, namely WordNet, to generate an enhanced ontology named WordNet-Plus. Third, the performance of WordNet-Plus is investigated under the ACR model. Due to the high coverage, WordNet-Plus can outperform WordNet on data sets having more fresh documents in classification. Fourth, extensive case studies of WordNet-Plus are given to demonstrate the rationality of WordNet-Plus construction. Quantitative evaluation is also conducted to further examine the quality of WordNet-Plus.

In the remainder of this paper, we first review the literature in Section 2. After the preliminary of ontology and the overview of ACR model are introduced in Section 3, two main components of ACR, namely, concept border generation and concept-based document representation, are presented in Sections 4 and 5 respectively. The technique details and the time complexity of ACR are presented in Section 6. The construction of WordNet-Plus is discussed in Section 7. Then, the experiment design and results are given in Sections 8 and 9. Finally, we conclude the paper.

2. Related work

As an important expert-edited ontology, WordNet has been used to improve the performance of clustering and classification. Hotho et al. [18,19] show that incorporating the synset and the hypernym as background knowledge into the document representation can improve the clustering results. Jing et al. [24] construct a term similarity matrix using WordNet to improve text clustering. However, their approach only uses synonyms and hyponyms, and fails to handle polysemy, and breaks the multi-word concepts into a group of single words. In Recupero's work [35], two strategies, namely, WordNet lexical categories (WLC) technique and WordNet ontology (WO) technique, are used to create a new vector space with low dimensionality for the documents. The authors in [39] successfully integrate the WordNet resource for document classification. They show improved classification results with respect to the Rocchio and Widrow-Hoff algorithms. A significant difference between our ACR model and the methods mentioned above is that we adopt a learning process to determine the dimensions in the new representation for the documents, which gives our method more adaptability in dealing with different document collections.

Wikipedia is an important online linguistic resource, which has been studied quite a lot for different purposes in recent year, such as clustering [21,20], classification [14,47,48], and semantic relatedness computation [15,51]. In clustering [21,20], the researchers

[2] A hierarchical directed acyclic graph is a directed acyclic graph with the layer information on each node. The head node of an edge must have a higher layer than the tail of the edge.