Expert Systems with Applications 36 (2009) 9095–9104

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity measures

Wei Song *, Cheng Hua Li, Soon Cheol Park

Department of Electronics and Information Engineering, Chonbuk National University, Jeonju, Jeonbuk 561-756, Republic of Korea

ARTICLE INFO

Keywords:
Genetic algorithm
Text clustering
Ontology
Wordnet
Latent semantic indexing

ABSTRACT

This paper proposes a self-organized genetic algorithm for text clustering based on ontology method. The common problem in the fields of text clustering is that the document is represented as a bag of words, while the conceptual similarity is ignored. We take advantage of thesaurus-based and corpus-based ontology to overcome this problem. However, the traditional corpus-based method is rather difficult to tackle. A transformed latent semantic indexing (LSI) model which can appropriately capture the associated semantic similarity is proposed and demonstrated as corpus-based ontology in this article. To investigate how ontology methods could be used effectively in text clustering, two hybrid strategies using various similarity measures are implemented. Experiments results show that our method of genetic algorithm in conjunction with the ontology strategy, the combination of the transformed LSI-based measure with the thesaurus-based measure, apparently outperforms that with traditional similarity measures. Our clustering algorithm also efficiently enhances the performance in comparison with standard GA and *k*-means in the same similarity environments.

1. Introduction

With the abundance of text documents available on the internet, the automatic partition of texts into previously unseen categories ranks top on the priority list for Information Retrieval (IR), and Pattern Recognition. However, the characteristics of polysemy and synonymy that exist in words of natural language have always been a challenge in the fields of IR and data mining. In many cases, humans have little difficulty in determining the intended meaning of an ambiguous word, while it is extremely difficult to replicate this process computationally. One main reason for this is that the existing retrieval solutions only relate documents that use identical terminology, while they ignore conceptual similarity of terms.

To address this problem, clustering algorithm is introduced first. Clustering is a popular unsupervised classification technique which groups the input space into *K* regions based on some similarity or dissimilarity metric. The partition is done such that patterns within a group are more similar to each other than patterns belonging to different groups (Frigui & Krishnapuram, 1999; Koontz, Narendra, & Fukunaga, 1975a, 1975b). Clustering is run-timely formed during the partition process, instead of being pre-defined as in case of text categorization, which commonly refers to the supervised partitioning of documents to "labeled" sets (Xia, Wang, & Yoshida, 2006). The task of documents clustering is both difficult and intensively studied in literature. A branch and bound algorithm uses a tree search technique to search the entire solution space (Koontz et al., 1975a, 1975b). It employs a criterion of eliminating sub trees which do not contain the optimal result. In this scheme, the number of nodes to be searched becomes huge as the size of the dataset becomes large. *k*-Means algorithm, one of the most widely used, attempts to solve the clustering problem into a fixed number of clusters *K* known in advance (Selim & Ismail, 1984). It is an iterative hill-climbing algorithm and solution suffering from the limitation of the sub-optimal which is known to depend on the choice of initial clustering distribution. Since stochastic optimization approaches can avoid convergence to a local optimization, these approaches can be used to find a globally optimal solution. Genetic algorithm (GA) belongs to the search techniques that mimic the principle of natural selection and heredity. It performs search in complex, large and multimode landscapes, and provides near-optimal solutions for objective or fitness function (Bandyopadhyay, Pal, & Aruna, 2004; Maulik & Bandyopadhyay, 2000). However, most of these clustering algorithms solely adopt vector space model (VSM) to represent text. That is, each unique term in vocabulary represents one dimension in feature space. The bag of words representation used for these clustering methods is often unsatisfactory as it ignores relationships between important terms which do not co-occur literally. This is due to the nature of text, where the same concept can be

* Corresponding author. Tel.: +82 63 270 2467; fax: +82 63 270 2461.
E-mail address: songwei9988@yahoo.com.cn (W. Song).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.12.046

represented by many different words, and words can have ambiguous meaning. Meanwhile, with the direct representation of text, there is a lack of more general concepts which can help identifying related topics. For example, a document about "canine" may not be related to a document about "feline" by the traditional clustering algorithms if there are only "canine" and "feline" in the different vectors. But if we add a more general concept "carnivore" to both documents, their semantic relationship is revealed. Thus, it is essential that a document clustering algorithm is regarded as a data clustering method combined with an appropriate document similarity measure.

In this paper we propose a modified genetic algorithm based on ontology for text clustering. We take advantage of thesaurus-based ontology and corpus-based ontology to provide a more accurate assessment of the similarity between documents. The lexical taxonomy Wordnet is designed in a tree-like hierarchical structure going from many specific terms at the lower levels to a few generic terms at the top (Hotho, Staab, & Stumme, 2003; Miller, 1995). We can use its hierarchical structure and broad-coverage taxonomy as thesaurus-based ontology. Meanwhile, a novel transform from the original latent semantic indexing (LSI) is proposed and demonstrated as the corpus-based ontology which can appropriately depict the associative semantic relationship in this study.

A variable string length GA using gene index to encode chromosome is developed to achieve the proper number of clusters. Meanwhile, considering the influence between the diversity of the population and the selective pressure, a self-organized evolution process is put forward in this article.

In the next section we give a brief review of ontology-based semantic similarity, and describe how we use it to compute in Wordnet. In Section 3 a transformed LSI model is proposed for corpus-based text representation, which is then used in conjunction with the thesaurus-based method as a hybrid strategy to evaluate document similarity measure. The details of genetic algorithm for text clustering based the ontology are described in Section 4. Experiment results are given in Section 5. Conclusions and future works are given in Section 6.

2. Ontology-based semantic similarity

Semantic similarity is a generic issue in the variety of application areas of Artificial Intelligence (AI) and Natural Language Processing (NLP). Similarity between two words is often represented by similarity between the concepts related with the two words. A number of semantic similarity methods have been developed in literature. Various similarity methods have proven to be useful in some specific application (Hotho & Maedche, 2001; Rada, Mili, Bichnell, & Blettner, 1989). In general, the semantic similarity measures can be categorized into two groups: edge-counting-based (or dictionary/thesaurus-based) methods and information theory-based (or corpus-based) methods (Li, Bandar, & Mclean, 2003). Assuming a lexical taxonomy is constructed in a tree like hierarchy with a node for a concept, it has proven that the minimum number of edges connecting concepts c_1 and c_2 is a metric for measuring the conceptual distance of c_1 and c_2 (Rada et al., 1989). The edge counting is useful for specific application with highly constrained taxonomies. However, lexical taxonomy may have irregular densities due to its broad coverage domain. Such a problem of nonuniformity can be corrected by the utilization of depth in the hierarchy where the word is found (Jiang & Conrath, 1997).

The basic idea of information theory-based methods is to define the similarity between two concepts as the maximum of the information context of the concept that subsumes them in the taxonomic hierarchy (Resnik, 1995). The information content of a concept depends on the probability of encountering an instance of the concept in a corpus. That is, the concept probability is obtained by the frequency of occurrence of the concept and its sub concept in the corpus. However, this method is rather difficult to tackle because of the limitation of the specific corpus application and its high computational complexity.

2.1. The semantic similarity in Wordnet

Wordnet is an online semantic dictionary which is developed at Princeton by a group led by Miller (1995). The version utilized in this paper is WordNet 2.0, which has 144,684 words and 109,377 synonym sets, named synsets. Wordnet organizes the lexicon by nouns, verbs, adjectives, and adverbs. Nouns, verbs, adjectives, and adverbs are represented by synsets. The synset reflects a concept in which all words have similar meaning. So it is interchangeable for synset by some syntax. The functions of synset include the concept definition for each word and the semantic relationship pointed to other related synsets. Wordnet provides a number of 18 kinds of relations to represent nouns concepts. The "ISA" hierarchical structure of the knowledge base is important in determining the semantic distance between words. Fig. 1 shows a part of such a hierarchical semantic knowledge base.

Given two concepts c_1 and c_2, the semantic similarity of $s(c_1, c_2)$ can be obtained from the tree-like hierarchical structure of Wordnet. One direct approach is to find the minimum length of the path connecting these two concepts. Fig. 1 shows a part of such a hierarchical semantic knowledge base. The shortest path between "teacher" and "student" is "teacher–educator–professional–adult–person–intellectual–student". The minimal length of the path is 6. The synset "person" is called the subsumer for concepts "teacher" and "student". While the minimal path length between "teacher" and "parent" is 9. Thus we could conclude that "student" is more similar to "teacher" than "parent" to "teacher". If a word has multiple meaning, various paths may exist. However, only the shortest path for semantic similarity calculation may be not so accurate. For example, the minimal length between "student" and "vehicle" is 6, less than from "student" to "parent". Whereas we could not conclude "student" is

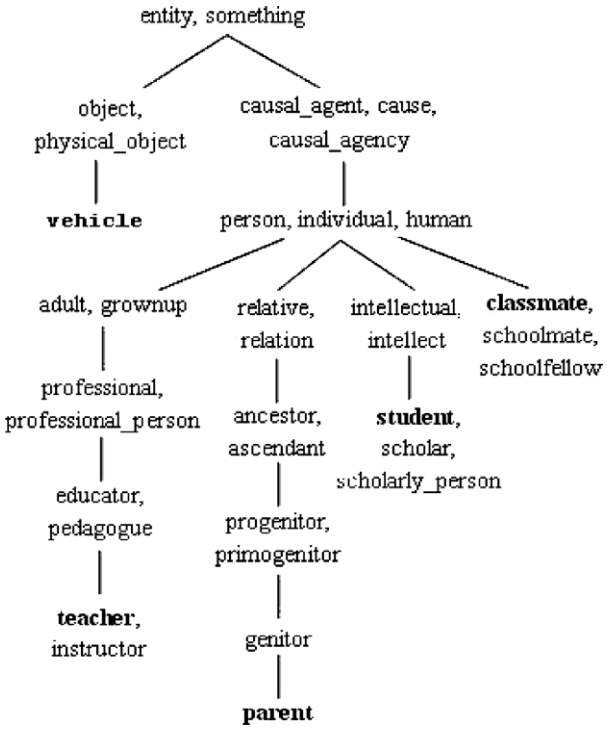

Fig. 1. A part of hierarchical semantic knowledge base.