Information Processing and Management 47 (2011) 287–299

Contents lists available at ScienceDirect

Information Processing and Management

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

Using a new relational concept to improve the clustering performance of search engines ☆

Lin-Chih Chen *

Department of Information Management, National Dong Hwa University, No. 1, Sec. 2, Da Hsueh Road, Shou-Feng, Hualien 97401, Taiwan

ARTICLE INFO

Article history:
Received 8 April 2008
Received in revised form 17 November 2008
Accepted 14 April 2010
Available online 7 May 2010

Keywords:
Document clustering
Semantic relation
Relational concept
Web search engines
Web documents

ABSTRACT

In this paper, we present a novel clustering algorithm to generate a number of candidate clusters from other web search results. The candidate clusters generate a connective relation among the clusters and the relation is semantic. Moreover, the algorithm also contains the following attractive properties: (1) it can be applied to multilingual web documents, (2) it improves the clustering performance of any search engine, (3) its unsupervised learning can automatically identify potentially relevant knowledge without using any corpus, and (4) clustering results are generated on the fly and fitted into search engines.

1. Introduction

Document clustering is an important technology that can automatically discover groups of similar web documents in a set of the web documents returned by a search engine (Berkhin, 2002). Organizing web documents by clustering can help the user get a sense of the major subject areas covered in the web document set and help the user find the relevant web documents more quickly. Clustering techniques can be divided into hierarchical and partitional methods (Berkhin, 2002; Cios, Pedrycz, Swiniarski, & Kurgan, 2007).

Hierarchical methods yield an entire sequence of nested partitions, and these methods can be either agglomerative or divisive. Agglomerative methods yield a sequence of nested partitions starting with one-document clusters recursively merging two or more of the most appropriate clusters. Divisive methods start with one cluster of all documents and recursively split the most suitable cluster. Many clustering algorithms belong to hierarchical methods, such as HAC (Voorhees, 1986), STC (Zamir & Etzioni, 1998), and DIVCLUS-T (Chavent, Lechevallier, & Briant, 2007).

Partitional methods find the clusters by partitioning the entire document collection into either a predetermined or an automatically derived number of clusters. Partitional methods can be considered an optimization procedure that tries to create high quality clusters according to a particular criterion function to achieve high intra-cluster similarity and low inter-cluster similarity. Many clustering algorithms belong to partitional methods, such as K-means (MacQueen, 1967), SRE (Zha, He, Ding, Simon, & Gu, 2001), and k-Attractors (Kanellopoulos, Antonellis, Tjortjis, & Makris, 2007).

In this paper, we propose a novel clustering algorithm, called *On-The-Fly Document Clustering* (OTFDC). OTFDC is based on a connective semantic relation between the clusters (as defined as Section 3), and it has the following four properties: (1) it can be applied to multilingual web documents, (2) it improves the clustering performance of any search engine, (3) it uses

☆ The experimental search engines used in this paper is available at http://cayley.sytes.net/english and http://cayley.sytes.net/chinese.

* Tel.: +886 3 863 3111; fax: +886 3 863 3100.
E-mail address: lcchen@mail.ndhu.edu.tw

0306-4573/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2010.04.001

unsupervised learning to automatically identify potentially relevant knowledge without using any corpus, and (4) it provides clustering results generated on the fly and fitted into search engines.

This paper is organized as follows. Section 2 introduces related work in the area of document clustering. Section 3 discusses OTFDC in detail. We then present some preliminary simulation results in Section 4. Finally, Section 5 concludes the paper.

2. Related work

This section reviews some important literature related to the paper. First, we provide a literature review on different concepts of the document clustering. Second, we discuss some online academic document clustering systems. Third, we provide a short description of Search engine Vector Voting (SVV).

2.1. Different concepts of the document clustering

Document clustering is an important data exploration technique that groups similar documents into a cluster. However, document clustering is a difficult problem since the documents contain large amounts of unstructured text (Szczepaniak, Segovia, Kacprzyk, & Zadeh, 2003). Many researchers use different concepts and methods to solve the clustering problem.

Some researchers have used the fuzzy concept to solve the clustering problem. Tjhi and Chen (2007) combined the possibilistic and fuzzy formulations of co-clustering for automatic categorization of large document collections. The combined approach allows the word memberships to represent fuzzy word partitions, however it captures the natural word clusters structure. Saraçoğlu, Tütüncü, and Allahverdi (2007) designed a two-layer structure and let the documents pass through it to find the similarities. In two-layer structures, they used predefined fuzzy clusters to extract feature vectors of related documents. The similarity measure is estimated based on these feature vectors.

Other researchers have used the domain concept to solve the clustering problem. Kerne, Koh, Sundaram, and Mistrot (2005) presented an iterative method for generative semantic clustering of related terms in the documents. Their method for generating semantic clustering is based on a quantitative model that represents mutual information between each new domain and the domains already in the document. Dai, Xue, Yang, and Yu (2007) presented a co-clustering based method to learn an unlabeled data set in an unknown domain. The learning process starts with a labeled data set from a known domain which is called an in-domain, and applies the learning knowledge to a related but different domain which is called an out-of-domain.

Further researchers have used the query concept to solve the clustering problem. Chang, Kim, and Raghavan (2006) constructed the query concepts to express the user's information needs, rather than trying to reformulate the initial queries. To construct the query concepts, they extracted all document features from each document and then clustered these document features into primitive concepts that are used to form query concepts. Li, Chung, and Holt (2008) used the query sequence to rearrange clustering results. They claimed a query sequence is frequent if it occurs in more than a certain percentage of the documents in all document sets.

2.2. Online document clustering systems

Several online document clustering systems have been proposed in literature. WebCat (Giannotti, Nanni, Pedreschi, & Samaritani, 2003) uses Transactional K-Means to generate desirable clustering results. CIIRarchies (Lawrie & Croft, 2003) uses a probabilistic technique to extract sentences and builds the cluster labels in a hierarchy via a recursive algorithm. Lingo (Osinski & Weiss, 2004) uses a combination of phrases and Singular Value Decomposition (SVD) on a term-document matrix to find meaningful long cluster labels. SHOC (Zhang & Dong, 2004) uses the suffix array for sentences extraction and presents the cluster labels in a hierarchy via Suffix Tree Clustering (STC). FIHC (Fung, Wang, & Ester, 2003) uses the frequent itemset-like approach to produce a hierarchical topic tree for clusters. SnakeT (Ferragina & Guli, 2008) also uses an approach similar to the frequent itemset to extract meaningful labels and builds the cluster labels with the bottom-up hierarchy construction process. But none of these systems are available on the Internet.

On the other hand, many of online document clustering systems are available on the Internet. Credo (Carpineto & Romano, 2004a, 2004b) uses a Formal Concept Analysis (FCA) to construct the cluster label in a hierarchy form that allows the user to query documents and see retrieval results organized in a browsable concept lattice. However, the results of Credo are not always a sentence form. Carrot2 (Osinski & Weiss, 2005; Weiss & Stefanowski, 2003) is an implementation of STC with extensions such as stop-words and stemming. It used sentences with variable length words as the cluster label, but these sentences were drawn as contiguous portions of the source document. CatS (M. Radovanović & Ivanović, 2006) uses a separate category tree derived from the dmoz Open Directory topic hierarchy to arrange all the cluster labels, but its results are restricted to a fixed set of topics. Highlight (Wu & Chen, 2003) adopts lexical analysis and a probabilistic technique to construct a concept hierarchy on the cluster label, but its classification is very rough. Vivisimo (Segev, Leshno, & Zviran, 2007) uses the Concise All Pairs Profiling (CAPP) to match all pairs of possible clusters in order to distinguish between any two clusters.