Clustering Web Pages into Hierarchical Categories

Zhongmei Yao, Louisiana Tech University, USA
Ben Choi, Louisiana Tech University, USA

ABSTRACT

Clustering is well suited for Web mining by automatically organizing Web pages into categories each of which contains Web pages having similar contents. However, one problem in clustering is the lack of general methods to automatically determine the number of categories or clusters. For the Web domain, until now there is no such a method suitable for Web page clustering. To address this problem, we discovered a constant factor that characterizes the Web domain, based on which we propose a new method for automatically determining the number of clusters in Web page datasets. We also propose a new Bidirectional Hierarchical Clustering algorithm, which arranges individual Web pages into clusters and then arranges the clusters into larger clusters and so on until the average inter-cluster similarity approaches the constant factor. Having the new constant factor together with the new algorithm, we have developed a clustering system suitable for mining the Web.

Keywords: information retrieval; knowledge classification; knowledge discovery; Semantic Web; Web mining

INTRODUCTION

We are interested in cluster analysis that can be used to organize Web pages into clusters based on their contents or genres (Choi & Yao, 2005). Clustering is an unsupervised discovery process for partitioning a set of data into clusters such that data in the same cluster is more similar to one another than data in other clusters (Berkhin, 2002; Everitt et al., 2001; Jain & Dubes, 1998; Jain et al., 1999). Typical application areas for clustering include artificial intelligence, biology, data mining, information retrieval, image processing, marketing, pattern recognition, and statistics (Berkhin, 2002; Everitt et al., 2001; Jain et al., 1999). Compared to
classification methods, cluster analysis has the advantage that it does not require any training data (i.e., the labeled data), but can achieve the same goal in that it can classify similar Web pages into groups.

The major aspects of the clustering problem for organizing Web pages are: To find the number of clusters \( k \) in a Web page dataset, and to assign Web pages accurately to their clusters. Much work (Agrawal et al., 1998; Dhillon et al., 2001; Ester et al., 1996; Guha et al., 1998a; Guha et al., 1998b; Hinneburg & Keim, 1999; Karypis & Kumar, 1999; Ng & Han, 1994; Rajaraman & Pan, 2000; Sander et al., 1998; Tantrum et al., 2002; Yao & Karypis, 2001; Zhang et al., 1996; Zhao & Karypis, 1999) has been done to improve the accuracy of assigning data to clusters in different domains, whereas no satisfactory method has been found to estimate \( k \) in a dataset (Dudoit & Fridlyand, 2002; Strehl, 2002) though many methods were proposed (Davies & Bouldin, 1979; Dudoit & Fridland, 2002; Milligan & Cooper, 1985). As a matter of fact, finding \( k \) in a dataset is still a challenge in cluster analysis (Strehl, 2002). Almost all existing work in this area assumes that \( k \) is known for clustering a dataset (e.g., Karypis et al., 1999; Zhao & Karypis, 1999). However in

Figure 1. The hierarchical structure produced for dataset DS1. Each box in this figure represents a cluster. The format of the description of a cluster is: its top three descriptive terms followed by (#docs, purity, entropy). Only the descriptions of clusters at the top level contain the \( F_1 \) scores.
Related Content

A Novel Approach for Detecting and Classifying Breast Cancer in Mammogram Images
[www.igi-global.com/article/novel-approach-detecting-classifying-breast/75544?camid=4v1a](www.igi-global.com/article/novel-approach-detecting-classifying-breast/75544?camid=4v1a)

A Modified Watershed Segmentation Method to Segment Renal Calculi in Ultrasound Kidney Images
[www.igi-global.com/article/modified-watershed-segmentation-method-segment/63351?camid=4v1a](www.igi-global.com/article/modified-watershed-segmentation-method-segment/63351?camid=4v1a)

Comparative Analysis of Artificial Neural Networks and Neuro-Fuzzy Models for Multicriteria Demand Forecasting
[www.igi-global.com/article/comparative-analysis-artificial-neural-networks/76296?camid=4v1a](www.igi-global.com/article/comparative-analysis-artificial-neural-networks/76296?camid=4v1a)
Computational Methods in Biomedical Imaging
www.igi-global.com/chapter/computational-methods-biomedical-imaging/10274?camid=4v1a