Applied Soft Computing 13 (2013) 959–966

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Probability based document clustering and image clustering using content-based image retrieval

M. Karthikeyan*, P. Aruna

Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Chidambaram, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 28 September 2010
Received in revised form 6 August 2012
Accepted 18 September 2012
Available online 3 October 2012

Keywords:
Document clustering
Word frequency
Content-based image retrieval
Major colour set
Global colour signature
Distribution block signature
Hue saturation value
Region of Interest
RGB histogram-based image retrieval

ABSTRACT

Clustering of related or similar objects has long been regarded as a potentially useful contribution of helping users to navigate an information space such as a document collection. Many clustering algorithms and techniques have been developed and implemented but as the sizes of document collections have grown these techniques have not been scaled to large collections because of their computational overhead. To solve this problem, the proposed system concentrates on an interactive text clustering methodology, probability based topic oriented and semi-supervised document clustering. Recently, as web and various documents contain both text and large number of images, the proposed system concentrates on content-based image retrieval (CBIR) for image clustering to give additional effect to the document clustering approach. It suggests two kinds of indexing keys, major colour sets (MCS) and distribution block signature (DBS) to prune away the irrelevant images to given query image. Major colour sets are related with colour information while distribution block signatures are related with spatial information. After successively applying these filters to a large database, only small amount of high potential candidates that are somewhat similar to that of query image are identified. Then, the system uses quad modelling method (QM) to set the initial weight of two-dimensional cells in query image according to each major colour and retrieve more similar images through similarity association function associated with the weights. The proposed system evaluates the system efficiency by implementing and testing the clustering results with *Dbscan* and *K-means* clustering algorithms. Experiment shows that the proposed document clustering algorithm performs with an average efficiency of 94.4% for various document categories.

1. Introduction

With the rapid development of information technology, the number of electronic documents and digital content of documents exceeds the capacity of manual control and management. People are increasingly required to handle wide ranges of information from multiple sources [1]. As a result, document clustering techniques are implemented by enterprises and organizations to manage their information and knowledge more effectively. Document clustering can be defined as the task of learning methods for categorizing electronic documents into their automatically annotated classes based on its contents [2]. It is widely applicable in areas such as search engines, web mining, information retrieval and topological analysis. Document clustering is a critical component of research in text mining. Traditional document clustering includes: (a) extracting feature vector of a document and (b) clustering of documents by parameters including similarity threshold and number of clusters, etc. Traditional document clustering, however, is the unsupervised learning; it cannot effectively group documents under the need of the user [3]. So, the proposed system concentrates on probability based topic oriented and semi-supervised document clustering approach.

Recently, as web and various documents contain a large number of images and text, it is necessary to cluster the images also. Content-based image retrieval (CBIR) applications are greatly needed for these applications [4]. With the increased emphasis on multimedia applications, the production of large information has resulted in a large volume of images that need to be properly indexed for retrieval in the future. Literature reports various techniques for CBIR and the most commonly utilized features are colour, shape and texture. The proposed system concentrates on image clustering by using the content-based image retrieval system to give more meaning to the proposed probability based topic oriented and semi-supervised document clustering method.

The structure of this paper is as follows: Section 2 discusses some related research work regarding document clustering and content-based image retrieval. Section 3 describes how document

* Corresponding author. Tel.: +91 9443665646.
E-mail address: mkshkarthik@yahoo.co.in (M. Karthikeyan).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.09.013

clustering is done by the probability based topic oriented and semi-supervised document clustering algorithm. Section 4 provides an image clustering by using content-based image retrieval method. The experimental results are given in Section 5. Finally some conclusion and discussion are given in Section 6.

2. Related works

Document clustering is a powerful technique to detect topics and their relations for information browsing, analysis and organization. In recent studies, many new technologies are introduced. Tseng [5] proposed an algorithm for cluster labelling to create generic titles based on external resources such as wordNet. This method first extracts category-specific terms such as cluster descriptors, and then these descriptors are mapped to generic terms based on a hypernym search algorithm. Trappey et al. [1] developed a document classification and search methodology, based on a neural network technology, by extracting key phrases from the document set, by means of automatic text processing and determining the significance of key phrases according to their frequency in text. Hao et al. [6] proposed a novel hierarchical classification method that generalizes support for vector machine learning.

Aliguliyev [7] developed a method to show assignment weight to documents that improves clustering solution because a document clustering has been traditionally investigated as a means of improving the performance of search engines by pre-clustering the entire corpus. Gong et al. [8] proposed a validity index-based method of adaptive feature selection, incorporating a new text stream clustering algorithm. Saracoglu et al. [9] developed a method for finding similar documents that uses predefined fuzzy clusters to extract feature vectors of related documents. Similarity measure is based on these vectors, and in 2008, they proposed a new approach on search for similar documents with multiple categories using fuzzy clustering that uses fuzzy similarity classification method and multiple categories vector method. Aliguliyev [7] again proposed a technique for automatic text summarization. He proved that summarization result not only depends on optimized function, and but also depends on a similarity measure. Horng et al. [10] proposed a hierarchical fuzzy clustering decision tree for the classification problem with large number of classes and continuous attributes. Song et al. [11] developed a method that uses genetic algorithm for text clustering based on ontology and evaluating the validity of various semantic measures.

Karray and Kamel [12] proposed a new concept-based mining model that analyzes terms on the sentence, document and corpus levels. It can effectively discriminate between non-important terms with respect to sentence semantics. Chim and Deng [13] developed a method for efficient phrase-based document similarity for clustering documents. They used phrase-based document similarity to compute the pair wise similarities of documents based on suffix. Frolov et al. [14] introduced a neural-network-based algorithm for word clustering.

Image classification deals with the problem of identifying an image in large database. It is desirable to classify and categorize image content automatically. Liu et al. [15] developed a region-based retrieval system with high-level semantic learning. It supports both queries by keyword and queries by region of interest. Liu and Hua [16] proposed a new index structure and query processing technique to improve retrieval effectiveness and efficiency. Gosselin and Cord [17] provided an algorithm within a statistical framework to extend active learning for online content-based image retrieval. Li et al. [18] proposed a framework based on multi-label neighborhood propagation for region-based image retrieval.

Pradhan and Prabhakaran [19] proposed an efficient indexing approach for 3-D human motion capture data, supporting queries involving both sub-body motions as well as whole-body motions. Aptoula and Lefevre [20] presented two morphology-based approaches, one making use of granulometries, independently computed for each sub quantized colour and another employing the principle of multi resolution histograms for describing colour, using morphological levelling and watersheds. Zhang and Ye [21] proposed new scheme to handle the noisy positive examples, by incorporating the methods of data cleaning and noise tolerant classifier.

3. System overview of probability based semi-supervised document clustering

Probability based topic oriented and semi-supervised document clustering is defined as follows: given a set S of n documents and a set T of k topics, the proposed system likes to partition the documents into k subsets $S_1, S_2, \ldots, S_k$, each corresponding to one of the topics, such that (i) the documents assigned to each subset are more similar to each other than the documents assigned to different subsets, and (ii) the documents of each subset are more similar to its corresponding topic than the rest of the topics. The functional components and data flow of proposed probability based topic oriented and semi-supervised document clustering method and image clustering using content-based image retrieval are depicted in Fig. 1. The proposed method concentrates also on image clustering by adapting the CBIR method. Literature reports various techniques for CBIR and the most commonly utilized features are colour, shape and texture.

The major steps involved in the proposed system are given below:

1. Documents of various categories are collected and stored in the database.
2. All the words and the images which appear in the documents are extracted and stored in a separate words database and image database with their corresponding categories.
3. From the words database distinct words are identified and probability is calculated.
4. During the clustering process, based on the higher probability of words, the documents are classified and clustered.
5. From the image database, the similar images are retrieved based on the given query Image by using content-based image retrieval system.
6. During image clustering, from the database image, global colour signature and distribution block signature are extracted.
7. For the query image, major colour set and distribution block signature are extracted.
8. MCS and DBS of the query image are compared with GCS and DBS of the database images, based on the similarity, and image clustering is done.
9. Finally, results are analyzed and compared with the results obtained by the existing algorithms *Dbscan* and *K-means* for document clustering and with RGB histogram-based image retrieval method for image clustering.

3.1. Document clustering by probability based topic oriented and semi-supervised clustering algorithm

The proposed new document clustering method, groups the documents according to the user's need. The main steps include: (1) design a multiple-attributes topic structure to represent user's need. (2) Make topic-semantic annotation for each document, and then compute topic-semantic similarity between documents. (3)