Automation in Construction 42 (2014) 36–49

Contents lists available at ScienceDirect

Automation in Construction

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

Automatic clustering of construction project documents based on textual similarity

Mohammed Al Qady *, Amr Kandil [1]

School of Civil Engineering, Purdue University, West Lafayette, IN 47907-2051, United States

ARTICLE INFO

Article history:
Received 7 July 2013
Received in revised form 4 February 2014
Accepted 8 February 2014
Available online 15 March 2014

Keywords:
Document management
Single pass clustering
Supervised/unsupervised learning methods

ABSTRACT

Text classifiers, as supervised learning methods, require a comprehensive training set that covers all classes in order to classify new instances. This limits the use of text classifiers for organizing construction project documents since it is not guaranteed that sufficient samples are available for all possible document categories. To overcome the restriction imposed by the all-inclusive requirement, an unsupervised learning method was used to automatically cluster documents together based on textual similarities. Repeated evaluations using different randomizations of the dataset revealed a region of threshold/dimensionality values of consistently high precision values and average recall values. Accordingly, a hybrid approach was proposed which initially uses an unsupervised method to develop core clusters and then trains a text classifier on the core clusters to classify outlier documents in a consequent refinement step. Evaluation of the hybrid approach demonstrated a significant improvement in recall values, resulting in an overall increase in F-measure scores.

1. Introduction

Automatic classification of documents as a supervised learning method requires a set of class labels and samples of each class in order to conduct the learning process before being able to perform predictions for new document instances. Usually, the classification procedure assumes that the classes are all inclusive (that they make a complete set of all the possible outcomes for any new instance) and that they are mutually exclusive (any new instance can belong to one and only one class). Where classes are static and predefined, the use of text classifiers for automatically organizing documents is appropriate. Documents are traditionally organized in construction projects according to fixed, abstract categories based on document metadata [1]. Examples of studies investigating the use of automatic text classification of construction documents include identifying the corresponding project division for minutes of meeting items [2] and classifying product documents to their relevant division in a construction information classification system [3].

While traditional methods of organizing construction project documents are simple and easy to use, they are not very useful for information retrieval unless the information seeker has thorough knowledge of the document body [1]. Information regarding a researched knowledge topic is almost always distributed over multiple categories thus requiring understanding of document content, not just metadata, to determine relevancy of a document to the researched topic; a time-consuming task that entails the application of human semantic capabilities. Also, the above-mentioned restrictions that constrain the use of classifiers do not apply with unsupervised methods: unsupervised methods do not require previous identification of all possible classes nor are they trained from sample data. The objective of this study is to evaluate the performance of an unsupervised learning text analysis technique in organizing project documents into groups of semantically similar documents; each group defined by its relation to a specific searchable knowledge topic. It is hypothesized that textual similarity between project documents accurately reflects semantic relationships between the documents and, when applied in document management and information retrieval tasks, can achieve results comparable to what humans recognize using their semantic capabilities. In the next section, the text analysis technique used in the study is presented along with several of its applications in previous works. Then the methodology implemented for the evaluation is presented, followed by a detailed analysis of the results. The study is concluded with a summary of the main results and a discussion on practical uses and limitations of implementing the proposed technique.

* Corresponding author at: 2775 Windwood Dr. #178 Ann Arbor, MI 48105. Tel.: +1 217 4196419.

E-mail addresses: malqady23@gmail.com (M. Al Qady), akandil@purdue.edu (A. Kandil).

[1] Tel.: +1 765 494 2246.

http://dx.doi.org/10.1016/j.autcon.2014.02.006
0926-5805/© 2014 Elsevier B.V. All rights reserved.

2. Clustering

Research on clustering methods for information retrieval dates back to the second half of the twentieth century. The main objective of clustering is to provide structure to a large dataset by organizing similar data together thus facilitating search and retrieval tasks. Clustering methods can be categorized according to the structure they generate into flat clustering and hierarchical clustering [4]. With flat- or non-hierarchical-clustering, the dataset is divided into a number of subsets of highly similar elements, dissimilar from elements in other clusters, with no relationship between the different clusters. The main advantage of this simple structure is low computational complexity in comparison with the more sophisticated hierarchical clustering methods. With hierarchical clustering, a complex structure of nested clusters is produced of the dataset. This is either done using a bottom–up approach, in which clusters start as individual items and pairs of similar items are joined together to form clusters, which are then joined together in successive steps until a single hierarchy is formed of the complete dataset. This approach is called agglomerative hierarchical clustering and is more popular than the top–down approach where the whole dataset is considered one cluster and is successively broken down into pairwise clusters until the level of the individual items is reached (also referred to as divisive hierarchical clustering). Flat clustering techniques include *K*-means and single pass clustering, while agglomerative hierarchical clustering techniques include single-link, complete-link and group-average. In terms of the exclusivity of cluster membership, clustering algorithms can be divided into hard clustering and soft clustering algorithms. In the former, membership of the items is limited to only one cluster. In the latter, the degree of association of each item to each cluster formed is determined [5].

Clustering was used in applications in many fields including organizing patient data in the medical field, classification of species in biological taxonomies and studying census and survey responses [4]. Clustering has a wide range of applications for data management in civil engineering. In the field of structural system identification, Saitta et al. [6] used *K*-means clustering to narrow down the number of candidate structural models in order to identify the best model that reflects actual sensor measurements of a structure. Principal component analysis was used to enable visualization of the various possible model clusters based on the most relevant model parameters. Cheng and Teizer [7] implemented clustering to identify objects from point cloud data of a laser scanner in order to enhance visibility of tower crane operators for safer hoisting operations. The DBSCAN algorithm was used for clustering. Similar to single pass clustering, DBSCAN starts with a randomly selected data point and successively forms clusters based on two user defined parameters: maximum allowable distance from the chosen point and minimum cluster size.

In data-mining of databases, Ng et al. [8] used *K*-means clustering to automatically group similar facility condition assessment reports of university facilities to investigate the relationship between reported deficiencies and facility types. A qualitative evaluation was used to verify the results of the investigation. Raz et al. [9] investigated the use of multiple techniques, including clustering, for developing models of good quality truck weigh-in-motion traffic data in order to facilitate identification of data anomalies. Two clustering techniques were investigated, *K*-means and Rectmix—a soft clustering algorithm. Implementation of the proposed mechanism by a domain expert was used to evaluate the accuracy and usefulness of the mechanism.

Clustering techniques were applied for defects' detection from images in several studies including: detection of potential defective regions in wastewater pipelines [10] and detection of rust in steel bridges to support decisions regarding bridge painting activities [11]. In the former study, region-growing segmentation – an application of single pass clustering to image data – was implemented for detecting defects in pipes using image analysis. Evaluation was performed based on the comparison of the results of the proposed technique with the inspection reports of a certified inspector using the following metrics: accuracy, recall and false alarm rate. In the latter study, the researchers highlight the limitation of *K*-means clustering for detecting rust in grayscale images of bridge members, namely: irregular illumination of images, low-contrast images that obscure rust areas, and debris on bridge members that create noise in the image analysis.

Clustering was widely used in image and video identification/processing. Brilakis et al. [12] developed a framework for managing digital images of construction sites. The framework divides an image into clusters that represent different construction materials in the image and uses the cluster features to identify the material from a database of material signatures. Evaluation was performed by testing the correctness of identification of five different construction materials in terms of precision, recall and effectiveness. The researchers describe the high accuracy of the bottom–up clustering technique implemented in the method. In video image processing, several studies utilized clustering to develop a codebook – or dictionary – of actions and/or poses used for comparing, identifying and classifying motions of workers in a construction activity [13,14]. In both studies, *K*-means clustering was used to limit the multitude of possible actions into a fixed set of poses. For evaluation, a supervised learning algorithm was applied to classify the motions of workers on a test video based on the developed codebook, and performance was determined based on accuracy of classification.

Several observations are noted from the above review. The majority of the studies utilized a flat, hard clustering approach. Generally, the required application dictates the choice of an appropriate clustering method; e.g. when the number of resulting clusters is known – or can be reasonably inferred – *K*-means clustering is an appropriate method (as in the case of detecting dark colored defect areas in gray-scale images), when multiple associations are feasible, a soft clustering approach is warranted. For evaluation of a clustering method and validation of the outcome, expert review was used in a number of the studies. In [4], the authors note the difficulty of evaluating clustering methods, and report using the comparison between an outcome and the clusters developed by domain experts as a common method for measuring performance.

For the purpose of this study a flat, hard clustering approach is deemed appropriate, for the reasons explained in the Methodology section. Clustering functions on the same basic assumption as classification—that similar documents form clusters that do not overlap with other non-similar document clusters (also referred to as the contiguity hypothesis). However clustering aims at identifying such document clusters without any external help from previously labeled instances (thus the unsupervised nature of the method). This is usually executed in an iterative process in which a specific procedure is repeated until a predefined condition is satisfied. Two main flat clustering techniques are reviewed below.

2.1. *K*-means

In *K*-means clustering, a number of *K* centroids is defined by the user and all instances in the dataset are assigned to the closest centroid (determined by Euclidean distance or cosine similarity). Then, centroids of all *K* clusters are calculated according to this assignment resulting in new centroid positions. All instances in the dataset are re-assigned to the new centroids and this iterative process is continued until cluster centroids remain constant, implying that the optimal centroid positions are identified (those that minimize the distance between each instance in a specific cluster and the cluster's centroid).