Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 82 (2016) 28 – 34

Procedia
Computer Science

Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia

Finding similar documents using different clustering techniques

Sumayia Al-Anazi, Hind AlMahmoud, Isra Al-Turaiki*

Information Technology Department
College of Computer and Information Sciences
King Saud University
Riyadh 12372, Saudi Arabia

Abstract

Text clustering is an important application of data mining. It is concerned with grouping similar text documents together. In this paper, several models are built to cluster capstone project documents using three clustering techniques: *k-means*, *k-means fast*, and *k-medoids*. Our datatset is obtained from the library of the College of Computer and Information Sciences, King Saud University, Riyadh. Three similarity measure are tested: *cosine similarity*, *Jaccard similarity*, and *Correlation Coefficient*. The quality of the obtained models is evaluated and compared. The results indicate that the best performance is achieved using *k- means* and *k-medoids* combined with cosine similarity. We observe variation in the quality of clustering based on the evaluation measure used. In addition, as the value of *k* increases, the quality of the resulting cluster improves. Finally, we reveal the categories of graduation projects offered in the Information Technology department for female students.

Peer-review under responsibility of the Organizing Committee of SDMA2016

Keywords: clustering; k-means; k-medoids; text mining; data mining; cosine similarity

1. Introduction

Today, with the rapid advancements in technology we are able to accumulate huge amounts of data of different kinds. *Data mining* emerged as a field concerned with the extraction of useful knowledge from data[1]. Data mining techniques have been applied to solve a wide range of real-world problems. Clustering is an unsupervised data mining technique where the labels of data objects are unknown. It is the job of the clustering technique to identify the categorisation of data objects under examination. Clustering can be applied to different kinds of data including text. When dealing with textual data, objects can be documents, paragraphs, or words[2]. *Text clustering* refers to the process of grouping similar text documents together. The problem can be formulated as follows: given a set of documents it is required to divide them into multiple groups, such that documents in the same group are more similar to each other than to documents in other groups. There are many applications of text clustering including: document organisation and browsing, corpus summarisation, and document classification[3].

* Corresponding author. Tel.: +0-966-11805-2909.
E-mail address: sumaiah.j@gmail.com, hindsalmahmoud@gmail.com, ialturaiki@ksu.edu.sa

Peer-review under responsibility of the Organizing Committee of SDMA2016
doi:10.1016/j.procs.2016.04.005

Traditional clustering techniques can be extended to deal with textual data. However, there are many challenges in clustering textual data. The text is usually represented in high dimensional space even when it is actually small. Moreover, correlation between words appearing in the text needs to be considered in the clustering task. The variations in document sizes is another challenge that affects the representation. Thus, the normalisation of text representation is required[2].

In this paper, we use data mining techniques in order to cluster capstone projects in information technology. In particular, we study graduation projects offered in the Information Technology department (IT) for female students at the College of Computer and Information Sciences, King Saud University, Riyadh. The goal is to to reveal the areas that the department encourages students to work on. The results of the study will be beneficial to both students and decision makers. For students, clustering graduation projects will help them find previous projects related to their own project idea. The study will also help the administration make right decisions when approving project proposals. We apply and compare three clustering techniques: *k-means*[4], *k-means fast*[5], and *k-mediods*[6]. In addition, three similarity measures are used to form clusters: *cosine similarity*[7], *Jaccard similarity*, and *Correlation Coefficient*[1]. The goal of the comparison is to find the best combination of clustering technique and similarity measure and to study the effect of increasing the number of clusters, k.

The rest of the paper is organised as follows: In Section 2, we review some of the literature in the field of text clustering. Section 3, describes our dataset, the steps taken to prepare it for data mining, and the data mining techniques and the similarity measures used in our experiment. The cluster evaluation measures and our main findings are discussed in Section 4. Finally, our paper concludes in Section 5.

2. Literature Review

Text clustering is one of the important applications of data mining. In this section, we review some of the related work in this field. Luo et al.[3] used the concepts of document *neighbors* and *links* in order to enhance the performance of *k-means* and *bisecting k-means* clustering. Using a pairwise similarity function and a given similarity threshold, the neighbors of a document are the documents that are considered similar to it. A *link* between two documents is the number of common neighbors. The concepts were used in the selection of initial cluster centroids and in document similarity measuring. Experimental results using 13 datasets showed better performances as compared to the standard algorithms.

Bide and Shedge[8] proposed a clustering pipeline to improve the performance of *k-means* clustering. The authors adopted a divide-and-conquer approach to cluster documents in the *20 Newsgroup dataset*[9]. Documents were divided into groups where preprocessing, feature extraction, and *k-means* clustering were applied on each group. Document similarity was calculated using the cosine similarity measure. The proposed approach achieved better results as compared to standard *k-means* in terms of both cluster quality and execution time.

Mishra et al.[10] used *k-means* technique to cluster documents based on themes present in each one. The main assumption was that a document may deal with multiple topics. The proposed approach, called *inter-passage based clustering*, was applied to cluster document segments based on similarity. After segments were preprocessed, keywords were identified for each segment using *term frequency-inverse document frequency*[11] and sentiment polarity scores[12]. Each segment was then represented using keywords and a segment score was calculated. Finally, *k-means* was applied to all segments. The resulting clusters showed high intra-cluster similarity and low inter-cluster similarity.

In general, algorithms used for clustering texts can be divided into: agglomerative, partitioning-based, and probabilistic-based algorithms[13]. Agglomerative algorithms iteratively merge documents into clusters based on pairwise similarity. The resulting clusters are organised into a cluster hierarchy (also *dendogram*). In partitioning algorithms, documents are split into mutually exclusive (non-hierarchical) clusters. The splitting process optimises the distance between documents within a cluster. Probabilistic clustering is based on building generative models for the documents. Partitioning algorithms for text clustering have been extensively studied in the literature. This is mainly due to the low computational requirements as compared to other clustering algorithms. In this paper, we choose to utilize three partitioning-based algorithms: *k-means*[4], *k-means fast*[5], and *k-mediods*[6] in order to cluster capstone projects.