Dynamic Document Clustering Using Singular Value Decomposition

Rashmi Nadubeediramesh, Department of Information Systems, University of Maryland Baltimore County (UMBC), Baltimore, MD, USA

Aryya Gangopadhyay, Department of Information Systems, University of Maryland Baltimore County (UMBC), Baltimore, MD, USA

ABSTRACT

Incremental document clustering is important in many applications, but particularly so in healthcare contexts where text data is found in abundance, ranging from published research in journals to day-to-day healthcare data such as discharge summaries and nursing notes. In such dynamic environments new documents are constantly added to the set of documents that have been used in the initial cluster formation. Hence it is important to be able to incrementally update the clusters at a low computational cost as new documents are added. In this paper the authors describe a novel, low cost approach for incremental document clustering. Their method is based on conducting singular value decomposition (SVD) incrementally. They dynamically fold in new documents into the existing term-document space and dynamically assign these new documents into pre-defined clusters based on intra-cluster similarity. This saves the cost of re-computing SVD on the entire document set every time updates occur. The authors also provide a way to retrieve documents based on different window sizes with high scalability and good clustering accuracy. They have tested their proposed method experimentally with 960 medical abstracts retrieved from the PubMed medical library. The authors’ incremental method is compared with the default situation where complete re-computation of SVD is done when new documents are added to the initial set of documents. The results show minor decreases in the quality of the cluster formation but much larger gains in computational throughput.

Keywords: Data Retrieval, Data Storage, Digital Libraries, Dimensionality Reduction Techniques, Dynamic Document Clustering, Grouping

INTRODUCTION

Educational institutions, industries, organizations and government agencies allocate substantial resources on research and development activities. Researchers around the world are working to devise methods and develop approaches to solve current problems, predict and prevent future concerns. However storing, updating, retrieving and grouping literature pertaining to various research topics has been a major concern. Much work has been done on dimensionality reduction techniques and information retrieval process. Digital libraries
must be capable of storing and delivering trillions of bytes of data to millions of users. Our goal is to address these key issues and provide an effective solution that results in reduction of storage space, retrieval time, memory usage and computation time and hence increase the overall performance. To achieve this goal we:

- Develop an incremental method for document clustering using singular value decomposition;
- Demonstrate the accuracy of the cluster formation and maintenance in an incremental manner;
- Empirically demonstrate the scalability of our proposed method using medical abstracts from Pubmed.

**Statement of the Problem**

Many public and private research organizations both domestically and internationally are working in collaboration to develop new and preventive vaccines, to investigate side effects and to detect outdated drugs and report its adverse effects. Health care professionals play a key role in providing patients and their families with timely and reliable information. Physicians practice evidence-based medicine by analyzing data available from research published in online journals. This helps them stay current and provide the best possible treatment for their patients. For example: Randomized-controlled studies provide evidence if a particular drug is effective in treating a particular patient population or not. Doctors use this information and apply it in their clinical practice to make a difference to their patients.

Patients often get information from a lot of resources, from family, friends, internet, and handouts at the doctor’s office. This helps them to understand the disease process, the etiology behind the disease and various treatment options available, and where they could find the best doctors that provide such advanced treatments. This is particularly true in case of rare genetic disorders with no or limited treatment options where patients can provide support to each other through support groups and share their experiences.

With the rapid growth in internet infrastructure and web technologies, medical text information is growing at an astounding pace. Storing the text data in its original form requires a high dimensional space. In addition updating, retrieving and grouping will be a cumbersome process. New research papers, articles and journals will be periodically published. Online databases should be designed to accompany these new manuscripts into the existing framework without huge modifications. Timely maintenance of these databases will increase storage space. For example: Removing obsolete documents and old documents that have low hit rates and downloads rates.

Although there are several document representation approaches, it needs a high dimensional space to represent the document and these approaches do not take into account the underlying semantic relationships Yu et al et al., (2008). In contrast, there are other matrix factorization techniques like called Singular Value Decomposition (SVD) that explicitly represent relations among very large number of terms and very large number of documents in which they occur. It can not only greatly reduce the dimensionality but also discover the important relationships between terms, between documents and between terms and documents Wang et al et al. (2009).

In this paper, the process of representing textual information in a matrix form is done by obtaining sparse term-document matrix (TDM) from text collections using Term Frequency and Inverse Document Frequency (TFIDF). TDM essentially represent the frequency of the terms with respect to the documents in which they occur. To capture the underlying semantics, robust correlation and to reduce the dimensionality $SVD$ is used.

$SVD$ is a feature transformation method (A method that projects original high dimensional space to lower dimension), that unveils the hidden data structure where terms and documents
Related Content

Early Warning from Car Warranty Data using a Fuzzy Logic Technique
www.irma-international.org/chapter/early-warning-car-warranty-data/38576/

Fuzzy Cluster Analysis of Larger Data Sets
www.irma-international.org/chapter/fuzzy-cluster-analysis-larger-data/38574/

Reinforcement Learning for Improving Gene Identification Accuracy by Combination of Gene-Finding Programs
www.irma-international.org/article/reinforcement-learning-improving-gene-identification/64643/

Information Therapy (Ix) Service and Patients’ Preference
www.irma-international.org/article/information-therapy-service-patients-preference/55492/

Research Commentary: Survival of the Fittest Algorithm or the Novelest Algorithm?
www.irma-international.org/article/research-commentary-survival-fittest-algorithm/51679/