A System To Provide Efficient Recommendation Based On Side Information Clustering- A Review

Nikita P.Katariya
Dept. of Computer Science & Engg
Priyadarshini Bhagwati College of Engineering
Nagpur, India
nikitakatariya@yahoo.com

Prof. M. S. Chaudhari
Dept. of Computer Science & Engg
Priyadarshini Bhagwati College of Engineering
Nagpur, India
Manojchaudhry2@gmail.com

***Abstract*—** Many data mining techniques have been proposed for mining useful patterns in text documents. Text mining is to research technologies to discover useful knowledge from enormous collections of documents, and to develop a system to provide knowledge and to support in decision making. Basically cluster means a group of similar data, document clustering means segregating the data into different groups of similar data. Clustering is a fundamental data analysis technique used for various applications such as biology, psychology, control and signal processing, information theory and mining technologies. Text mining is not a stand-alone task that human analysts typically engage in. The goal is to transform text composed of everyday language into a structured, database format. In many text mining applications, side-information is available along with the text documents. Such side-information may be of different kinds, such as document provenance information, the links in the document, user-access behavior from web logs, or other non-textual attributes which are embedded into the text document. Such attributes may contain a tremendous amount of information for clustering purposes. Therefore, principled way is needed to perform the mining process, so as to maximize the advantages from using this side information.

Keywords*-text mining; cluster, side-information*

*****

I. INTRODUCTION

Text mining refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling.

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted.

A. Side Information [1]

In many text mining applications, side information is available along with the text documents. Such information may be of different kinds such as document provenance information, the links in the document, user-access behavior from web logs, or other non textual attribute which are embedded into the text document. Some examples of such side-information are as follows:

- In an application user access behavior of web documents, the user-access behavior may be captured in the form of web logs can be tracked. For each document, the meta-information may correspond to the browsing behavior of the different users. Such logs can be used to enhance the quality of the mining process in a way which is more meaningful to the user, and also application-sensitive. This is because the logs can often pick up subtle correlations in content, which cannot be picked up by the raw text alone.
- Many text documents contain links among them, which can also be treated as attributes. Such links contain a lot of useful information for mining purposes. As in the previous case, such attributes may often provide insights about the correlations among documents in a way which may not be easily accessible from raw content.
- Many web documents have meta-data associated with them which correspond to different kinds of attributes such as the provenance or other information about the origin of the document. In other cases, data such as ownership, location, or even temporal information may be informative for mining purposes. In a number of network and user-sharing applications, documents may be associated with user-tags, which may also be quite informative.

II. LITERATURE SURVEY

Many methods for text classification have been proposed. A number of statistical classification and machine learning techniques has been applied to text categorization including regression models, nearest neighbor classifiers, decision trees, Bayesian classifiers and neural networks. All these methods are based on pure text data and do not work for the cases in which the text data is combined with other forms of data. The use of side information with the text can improve the efficiency of text classification. The side information may be of different types such as document provenance information, the links in the document, user access behavior from web logs, or other non textual information.

A. Text Mining [10]

Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. Text databases are rapidly growing due to the increasing amount of information available in electronic form, such as electronic publications, various kinds of electronic documents, e-mail, and the World Wide Web. Nowadays most of the information in government, industry, business, and other institutions are stored electronically, in the form of text databases. Data stored in most text databases are semi structured data in that they are neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, and category and so on, but also contain some largely unstructured text components, such as abstract and contents. There have been a great deal of studies on the modeling and implementation of semi structured data in recent database research. Moreover, information retrieval techniques, such as text indexing methods, have been developed to handle unstructured documents. Traditional information retrieval techniques become inadequate for the increasingly vast amounts of text data. Typically, only a small fraction of the many available documents will be relevant to a given individual user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. Thus, text mining has become an increasingly popular and essential theme in data mining.

B. Information Retrieval [9]

Information retrieval is a field that has been developing in parallel with database systems for many years. Unlike the field of database systems, which has focused on query and transaction processing of structured data, information retrieval is concerned with the organization and retrieval of information from a large number of text-based documents. Since information retrieval and database systems each handle different kinds of data, some database system problems are usually not present in information retrieval systems, such as concurrency control, recovery, transaction management, and update. Also, some common information retrieval problems are usually not encountered in traditional database systems, such as unstructured documents, approximate search based on keywords, and the notion of relevance. There exist many information retrieval systems, such as on-line library catalog systems, on-line document management systems, and the more recently developed Web search engines. A typical information retrieval problem is to locate relevant documents in a document collection based on a user's query, which is often some keywords describing an information need, although it could also be an example relevant document. In such a search problem, a user takes the initiative to "pull" the relevant information out from the collection; this is most appropriate when a user has some ad hoc information need, such as finding information to buy a used car. When a user has a long-term information need , a retrieval system may also take the initiative to "push" any newly arrived information item to a user if the item is judged as being relevant to the user's information need. Such an information access process is called information filtering, and the corresponding systems are often called filtering systems or recommender systems.

C. Clustering

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents, while maximizing inter-cluster distances. A distance measure thus lies at the heart of document clustering. Clustering is the most common form of unsupervised learning and this is the major difference between clustering and classification.

No super-vision means that there is no human expert who has assigned documents to classes. In clustering, it is the distribution and makeup of the data that will determine cluster membership. Clustering is sometimes erroneously referred to as automatic classification; however, this is inaccurate, since the clusters found are not known prior to processing whereas in case of classification the classes are pre-defined. In clustering, it is the distribution and the nature of data that will determine cluster membership, in opposition to the classification where the classifier learns the association between objects and classes from a so called training set, i.e. a set of data correctly labeled by hand, and then replicates the learnt behavior on unlabeled data.

D. Clustering Applications

Clustering is the most common form of unsupervised learning and is a major tool in a number of applications in many fields of business and science. Following are the basic directions in which clustering are used.

- **Finding Similar Documents**: This feature is often used when the user has spotted one "good" document in a search result and wants more-like-this. The interesting property here is that clustering is able to discover documents that are conceptually alike in contrast to search-based approaches that are only able to discover whether the documents share many of the same words.

- **Organizing Large Document Collections**: Document retrieval focuses on finding documents relevant to a particular query, but it fails to solve the problem of making sense of a large number of uncategorized documents. The challenge here is to organize these documents in a taxonomy identical to the one humans would create given enough time and use it as a browsing interface to the original collection of documents.
- **Duplicate Content Detection**: In many applications there is a need to find duplicates or near-duplicates in a large number of documents. Clustering is employed for plagiarism detection, grouping of related news stories and to reorder search results rankings. Note that in such applications the description of clusters is rarely needed.
- **Recommendation System**: In this application a user is recommended articles based on the articles the user has already read. Clustering of the articles makes it possible in real time and improves the quality a lot.
- **Search Optimization** : Clustering helps a lot in improving the quality and efficiency of search engines as the user query can be first compared to the clusters instead of comparing it directly to the documents and the search results can also be arranged easily

III. RELATED WORK

The problem of text-clustering has been studied widely by the database community. The major focus of the work has been on scalable clustering of multidimensional data of different types. Clustering problem has also been studied quite extensively in the context of text-data. A survey of text clustering methods may be found in [10]. One of the most well known techniques for text-clustering is the scatter-gather technique, which uses a combination of agglomerative and partitional clustering. An Expectation Maximization (EM) method for text clustering has been proposed in [11].

C.C.Aggarwal and C.X.Zhai [10] has done a survey of text clustering algorithm. The problem of clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks. The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function. The problem of clustering can be very useful in the text domain, where the objects to be clusters can be of different granularities such as documents, paragraphs, sentences or terms. Clustering is especially useful for organizing documents to improve retrieval and support browsing. The study of the clustering problem precedes its applicability to the text domain. Traditional methods for clustering have generally focused on the case of quantitative data, in which the attributes of the data are numeric.

Michael Steinbach, George Karypis and Vipin Kumar [3] have given the results of an experimental study of some common document clustering techniques. In particular, they have compared the two main approaches to document clustering, agglomerative hierarchical clustering and K-means. Hierarchical clustering is often portrayed as the better quality clustering approach, but is limited because of its quadratic time complexity. In contrast, K-means and its variants have a time complexity which is linear in the number of documents, but are thought to produce inferior clusters. Sometimes K-means and agglomerative hierarchical approaches are combined so as to "get the best of both worlds." However, results indicate that the bisecting K-means technique is better than the standard K-means approach and as good as or better than the hierarchical approaches that they tested for a variety of cluster evaluation metrics. They proposed an explanation for these results that is based on an analysis of the specifics of the clustering algorithms and the nature of document data.

Ning Zhong, Yuefeng Li, and Sheng-Tang Wu [4] have proposed a method for effective pattern discovery for text mining. Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase) based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. They presented an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

Shi Zhong [8] has proposed a method for efficient streaming text clustering. Clustering data streams has been a new research topic, recently emerged from many real data mining applications, and has attracted a lot of research attention. However, there is little work on clustering high-dimensional streaming text data. They proposed an efficient online spherical k-means (OSKM) algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams. The OSKM algorithm modifies the spherical k-means (SPKM) algorithm, using online update based on the well known Winner-Take-All competitive learning. It has been shown to be as efficient as SPKM, but much superior in clustering quality. The scalable clustering strategy was previously developed to deal with very large data bases that cannot fit into a limited memory and that are too expensive to read/scan multiple times. Using the strategy, one keeps only sufficient statistics for history data to retain the contribution of history data and to accommodate the limited memory. To make the proposed clustering algorithm adaptive to data streams, they introduced a forgetting factor that applies exponential decay to the importance of history data. The older a set of text documents, the less weight they carry. Their experimental results demonstrated the efficiency of the proposed algorithm and reveal an intuitive and an interesting fact for clustering text streams one need to forget to be adaptive.

R. Sagayam, S.Srinivasan and S. Roshni[9] has done a survey of text mining: retrieval, extraction and indexing Techniques. Text mining is the analysis of data contained in natural language text. Text mining works by transposing words and phrases in unstructured data into numerical values which can then be linked with structured data in a database and analyzed with traditional data mining techniques. Text information retrieval and data mining has thus become increasingly important.

IV. Problem Definition

The problem of text clustering arises in the context of many application domains such as the web, social networks, and other digital collections. The rapidly increasing amounts of text data in the context of these large online collections has led to an interest in creating scalable and effective mining algorithms. A tremendous amount of work has been done in recent years on the problem of clustering in text collections in the database and information retrieval communities. Primarily it has been designed for the problem of pure text clustering. In many application domains, a tremendous amount of side information is also associated along with the documents. This is because text documents typically occur in the context of a variety of applications in which there may be a large amount of other kinds of database attributes or meta information which may be useful to the clustering process. The need is to collect and create cluster of side information to improve the performance of search. The core of the approach is to determine a clustering technique in which the text attributes and side-information provide similar hints about the nature of the underlying clusters. The work will be extended to analyze the proposed solution and evaluate the performance using side information under different search parameter.

V. Proposed Work

The proposed work will be based on an approach for clustering text data with side information. The project will use the auxiliary information in order to provide additional insights, which can improve the quality of clustering. An approach will be designed in order to magnify the coherence between the text content and the side information. The work will be divided into five parts

1. Data set collection
2. Side information gathering
3. Information searching without side information
4. Information searching with side information
5. Analysis of results

The project will provide a first approach to using other kinds of attributes in conjunction with text clustering. Such an approach is especially useful, when the auxiliary information is highly informative, and provides effective guidance in creating more coherent clusters.

The main objective of this project is to design a recommendation search engine based on the side information clustering. The primary goal in this project is to study the clustering problem; such an approach can also be extended in principle to other data mining problems in which auxiliary information is available with text. The aim of the proposed system is to magnify the coherence between the text content and the side-information. Also in cases, in which the text content and side-information do not show coherent behavior for the clustering process, the effects of those portions of the side-information are marginalized.

VI. Conclusion

Text data clustering arises in the context of many application domains. The text data clustering with side information gives a recommendation search engine to improve the clustering process. The proposed algorithm can be easily applied to existing text mining system. The proposed system will show that the use of side information can greatly enhance the quality of text clustering and classification while maintaining a high level of efficiency.

References

[1] Charu C. Aggarwal, Fellow, IEEE Yuchen Zhao, and Philip S. Yu, Fellow, IEEE, "On the Use of Side Information for Mining Text Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 6, JUNE 2014. pp. 1415-1429

[2] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, "Scatter/Gather: A cluster-based approach to browsing large document collections," in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318–329.

[3] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.

[4] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012, pp. 30-44.

[5] C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in Mining Text Data. New York, NY, USA: Springer, 2012.

[6] I. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in Proc. ACM KDD Conf., New York, NY, USA, 2001, pp. 269–274

[7] I. Dhillon, S. Mallela, and D. Modha, "Information-theoretic coclustering," in Proc. ACM KDD Conf., New York, NY, USA, 2003, pp. 89–98.

[8] S. Zhong, "Efficient streaming text clustering," Neural Netw., vol. 18, no. 5–6, pp. 790–798, 2005.

[9] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", International Journal of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5, September 2012,pp.1443-1446.

[10] C. C. Aggarwal and C.-X. Zhai, "Mining Text Data", New York, NY, USA: Springer, 2012.

[11]T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering", in Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488–495.