On the Utilization Aspect of Document Data for Mining the Side Information

N.S. Krishna Prasad[1], S. Dhana Sekaran[2]

[1, 2] Rajalakshmi Engineering College, Chennai, Anna University, India

Abstract: ***In text mining applications, side-information is also available along with the text documents. This side-information can be like document provenance information, links existing inside the document, web logs based on user-access behavior, or non-textual attributes which exist in the text document. Such attributes will contain remarkable amount of information for clustering purposes. Usually it's difficult to estimate the importance of this side-information when they are noisy. In these scenarios, there is a huge amount of risk involved in incorporating this side-information into the mining process, since they can add noise to the process rather than improving the quality of the mining process. We need a standard way to perform the mining process, so that we make best use of the advantages based on this side information. In this paper, we propose an algorithm to create an effective clustering approach, based on the combination of traditional partitioning algorithms with probabilistic models. We also show how to illustrate methodology to the classification problem.***

Keywords: Data mining, clustering, Text documents, partitioning algorithm.

1. Introduction

Data mining is the process of analyzing data from various perceptions and briefing it into valuable information which can be used to maximize income, drill down the costs, or at times both. Data mining tool is used in analyzing data. In Data mining users can analyze data from various angles or dimensions, based on which it can be categorized and used to summarize the acknowledged relationships. Data mining is used to find the patterns among fields in bulky relational databases.

Data mining tools could be used to answer the business questions which were too time consuming to resolve. The name data mining is derived from the resemblances between identifying for valuable information in a bulky database and mining a mountain for a vein of valuable ore. In both of these processes we require either examining through an enormous amount of material, or logically examining it to find where the value resides.

Data mining is majorly used by organizations where there is a robust consumer focus - retail, financial, communication, and marketing organizations. It helps organizations to conclude the relationships among internal factors like skills of the staff's available, product price, positioning of the product and external factors like demographics of customers, economic indicators and competition. It also helps them to achieve the sales impact, satisfaction of customers, and profits of the corporate. At last, it allows them to narrow down into summary information to view detail transactional data.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

WalMart is pioneering massive data mining to transform its supplier relationships. WalMart captures point-of-sale transactions from over 2,900 stores in 6 countries and continuously transmits this data to its massive 7.5 terabyte Teradata data warehouse. WalMart allows more than 3,500 suppliers, to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, WalMart computers processed over 1 million complex data queries.

Data Mining is also known as knowledge discovery from data, Extraction of interesting patterns or knowledge from huge amount of data. Data mining — core of knowledge discovery process.

Data Mining consists of two tasks.
1) Predictive tasks
2) Descriptive tasks

Predictive tasks - Predict the value of the attribute based on the value of other attributes.

Descriptive tasks - To derive patterns that summarize the underlying relationship between data.

Steps involved in Data Mining are as follows:

Data Integration: Initially data will be collected and integrated from all the sources.

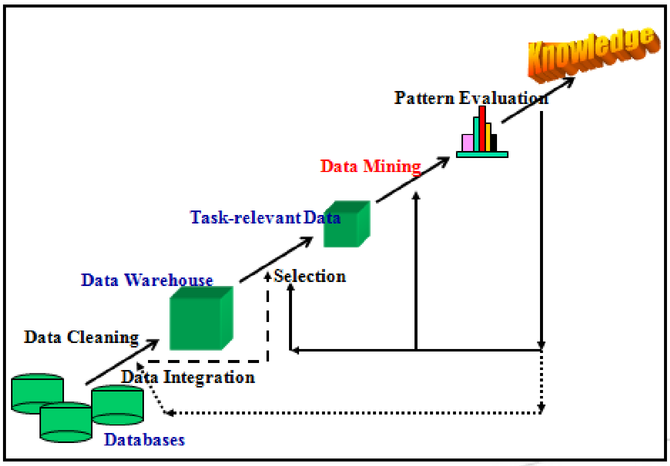

Figure 1: Knowledge Discovery Process

Data Cleaning: Data collected could contain errors, missing values, noisy or inconsistent data. So we need to apply various mechanisms to get rid of these irregularities.

Data Transformation: Normally cleaned data are not ready for mining since we are supposed to transform them into forms suitable for mining. The mechanisms used to achieve this are smoothing, aggregation, normalization.

Data Mining: Finally we are prepared to apply data mining techniques on the transformed data to identify the interesting patterns. Traditional techniques like clustering and association analysis are some of the popular mechanisms used for data mining.

Key techniques for Data Mining include Association, Classification, Clustering, Prediction, Sequential patterns, Decision trees, Combinations and Long-term (memory) processing.

Data Mining – Clusters
Cluster is a collection of Data objects. Similar to one another, within the same cluster and Dissimilar to the objects in other clusters. Cluster analysis –Used to find similarities between data as per the characteristics found in the data and grouping similar data objects into clusters. A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Clustering Approaches include Partitioning approach, Hierarchical approach, Density-based approach, Grid-based approach, Model-based, Frequent pattern-based and User-guided or constraint-based.

Data Selection: Usually we don't require all the data that we have collected in earlier step. So we need to filter the data which are useful for data mining.

1.1 Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

1.2 Clustering Approaches

a) **Partitioning Approach:** Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors. Typical methods: k-means, k-methods, CLARANS

b) **Hierarchical Approach:** Create a hierarchical decomposition of the set of data (or objects) using some criterion. Typical methods: Diana, Agnes, BIRCH, ROCK, CAMELEON

c) **Density-based Approach:** Based on connectivity and density functions. Typical methods: DBSACN, OPTICS, DenClue

d) **Grid-based Approach:** Based on a multiple-level granularity structure. Typical methods: STING, WaveCluster, CLIQUE

e) **Model-based**: A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other. Typical methods: EM, SOM, COBWEB

f) **Frequent pattern-based:** Based on the analysis of frequent patterns. Typical methods: pCluster

g) **User-guided or constraint-based:** Clustering by considering user-specified or application-specific constraints. Typical methods: COD (obstacles), constrained clustering.

2. Related Work

The major part of the project development sector considers and fully survey all the required needs for developing the project. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

In recent trends text clustering has become an issue due to enormous quantity of unstructured data which is existing in numerous forms like web, social networks and other information networks. In many cases, data is not just available in text form alone. There is a lot of side-information available along with the text documents. This side-information can be like document provenance information, links existing inside the document, web logs based on user-access behavior, or non-textual attributes which exist in the text document. Such attributes will contain remarkable amount of information for clustering purposes.

Usually it's difficult to estimate the importance of this side-information when they are noisy. In these scenarios, there is a huge amount of risk involved in incorporating this side-information into the mining process, since they can add noise to the process rather than improving the quality of the mining process. We need a standard way to perform the mining process, so that we make best use of the advantages based on this side information. In this paper, we propose an algorithm to create an effective clustering approach, based on the combination of traditional partitioning algorithms with probabilistic models. We also show how to illustrate a methodology to the classification problem.

In order to achieve this goal, we will combine a partitioning approach with a probabilistic estimation process, which determines the coherence of the side-attributes in the clustering process. A probabilistic model on the side information uses the partitioning information (from text attributes) for the purpose of estimating the coherence of different clusters with side attributes.

The partitioning approach is specifically designed to be very efficient for large data sets. This can be important in scenarios in which the data sets are very large. We will showcase the experimental results on a number of real data sets, and illustrate the effectiveness and efficiency of the approach presented a method for text clustering with the use of side-information.

In order to propose the clustering method, we have combined an iterative partitioning mechanism with a probability estimation process which calculates the significance of various kinds of approach pertaining to side-information. The results demonstrate the use of side-information can significantly improve the quality of text clustering.

In recent analysis on numeric data streams, problems of text and categorical data present various challenges due to large and un-ordered nature of corresponding attributes. So we recommend text and categorical data stream clustering algorithms. We also recommend approach for stream clustering based on condensation which summarizes the stream into a number of fine grained cluster droplets. These droplets could be used in conjunction with different user queries to form the clusters. So this proposes an online based analytical processing approach to stream clustering.

The problem of clustering text and categorical data streams. This problem is relevant in a number of web related applications such as news group segmentation, text crawling, and target marketing for electronic commerce. Some applications of text and categorical data stream clustering are many portals on the World Wide Web provide real time news and other articles which require quick summarization and filtering. Such methods often require effective and efficient methods for text segmentation

Many web crawlers continuously harvest thousands of web pages on the web, which is subsequently summarized by human effort. When the volume of such crawls is significant, it is not realistically possible to achieve this goal by human effort. In such applications, data stream clustering algorithms can be helpful in organizing the crawled resources into coherent sets of clusters.

In many electronic commerce applications, large volumes of transactions are processed on the World Wide Web. Such transactions can take the form of categorical or market basket records. In such cases, it is often useful to perform real time clustering for target marketing.

The algorithm was tested on a numerous synthetic and real data sets. We found the algorithm to be highly effective in being able to quickly adapt to variations in the data stream and recognize the underlying temporal locality.

Our method turns out to be much more effective, and the advantage was greater when the query was restricted to a particular user-specified horizon. We also tested the method for scalability, and it turns out to be highly efficient over a variety of data sets.

Based on real data mining applications, clustering data streams has attracted a lot of research attention. But clustering high dimensional streaming text date is not quite easy. To achieve this we need to combine the proficient online spherical k-means (OSKM) algorithm with existing scalable clustering strategy so that we can accomplish quick and adaptive clustering of text streams. OSKM algorithm is obtained by modifying the SPKM algorithm using online update based on Winner-Take-All competitive learning. OSKM has been proved to be as efficient as SPKM and also it provides superior clustering quality.

In order to use the proposed algorithm on data streams, we introduce a factor which applies exponential decay to the prominence of history data. Our result exhibits the effectiveness of the proposed algorithm and discloses instinctive and exciting information for clustering text streams.

There are several challenges being encountered: In real time high efficiency is required from every algorithm; data volume is enormous so that it cannot be stored in memory all at once; it's not advisable to perform repeated scans from secondary storage device which causes time delays and algorithms used in mining should be compatible with data patterns that keep changing over time.

Text clustering has become an increasingly important technique for unverified document organization, automatic extraction of topic and quick information or filtering. There has been enormous journalism on text clustering but only limited work on clustering streaming text data such as news streams.

3. Proposed Approach

To analyze and compare between the url and document, supporting links need to be analyzed. This will display, analyze & mine the corresponding paper & in addition also its related recent papers. Data mining techniques will be used to discover patterns from the web. In addition to mining the specific content related to side information we also analyse the corresponding web links.

Any group of words can be chosen as the stop words for a given purpose. For some search machines, these are some of the most common, short function words, such as the, is, at, which, and on. In this case, stop words can cause problems when searching for phrases that include them, particularly in names such as 'The Who', 'The', or 'Take That'. Stemming is the process for reducing derived words to their stem or base or root form— usually a written word form. It's not mandatory for the stem to be identical to the root of the word; generally related words map to the same stem, even if this stem is not in itself a valid root.
User feeds the input text document into the system for Mining. System identifies the stemming words and short function words. System then crawls through the web links and compares the document and its link pages then produces an efficient text mined data.

4. System Architecture

The major part of the project development sector considers and fully survey all the required needs for developing the project. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

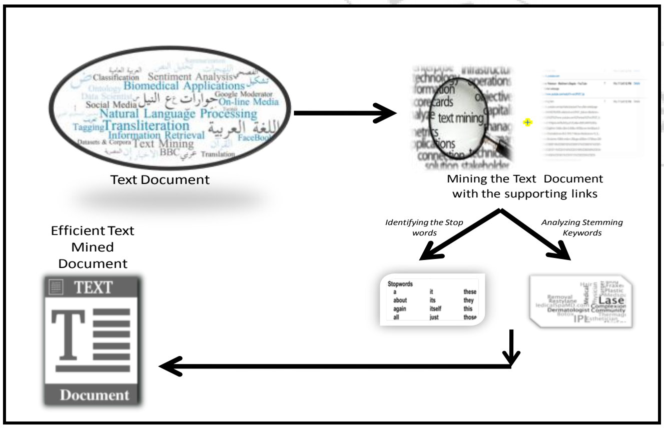
Text Document
Mining the Text Document with the supporting links
Identifying the Stop words
Analyzing Stemming Keywords
Efficient Text Mined Document
TEXT
Document

Figure 2: Architecture Diagram

Generally algorithms shows a result for exploring a single thing that is either be a performance, or speed, or accuracy, and so on. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. System architecture can comprise system components, the externally visible properties of those components, the relationships (e.g. the behavior) between them.

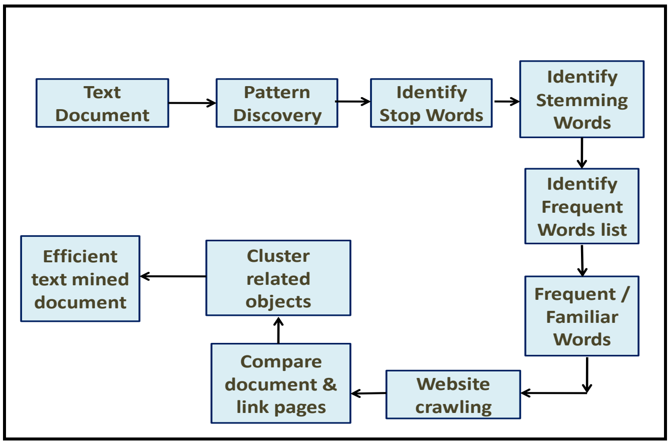
Text Document
Pattern Discovery
Identify Stop Words
Identify Stemming Words
Identify Frequent Words list
Frequent / Familiar Words
Website crawling
Compare document & link pages
Cluster related objects
Efficient text mined document

Figure 3: Architecture Diagram – Detailed view

5. Implementation

Following are the most frequently used project management

Methodologies in the project management practice:
1) Text Document
2) Mining Text and Supporting Links
3) Identify Steaming
4) Identify Stop words
5) Words Frequent List
6) Frequency List and Closed Pattern
7) Efficient Mined Document
8) Performance Evaluation

5.1 Text Document

The Text module is included in Drupal Core. When enabled, the Text module can be used to define simple text field types. The Text module defines various text field types for the Field module. A text field may contain plain text only, or optionally, may use Drupal's Text Filters to securely manage HTML output. Text input fields may be either a single line (text field), multiple lines (text area), or for greater input control, a select box, checkbox, or radio buttons. If desired, the field can be validated, so that it is limited to a set of allowed values.

5.2 Mining Text and Supporting Links

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. Each formula data refers to an Xti. The Xti specifies a particular supporting link record from the collection stored in the workbook. The Xti and supporting link record together specify where the data used by the formula element.

5.3 Identify Steaming

Words having similar logic interpretations can be considered as equivalent. This also reduces the dictionary size that is the number of distinct terms needed for representing a set of documents.

5.4 Identify Stop Words

Most Search Engines do not consider extremely common words in order to save disk space or to speed up search results. Stop words are words which are filtered out prior to, or after, processing of natural language data (text). Example of short functioning words as the, is, at, which, and on.

5.5 Words Frequent List

Word lists by frequency are lists of words grouped by frequency of occurrence within some given text either by levels or as a ranked list, serving the purpose of vocabulary acquisition.

5.6 Frequency List and Closed Pattern

Frequency list is a sorted list of words together with their frequency, where frequency here usually means the number of occurrences in a given text, from which the rank, less meaningful, can be derived. Closed pattern mining algorithm can be adapted to mine max pattern.

5.7 Efficient Mined Document

The conversion of words to concepts has been performed using a vocabulary and computational techniques. This will display, analyze & mine the corresponding paper & in addition also it's related recent papers.

5.8 Performance Evaluation

Here in this paper we use Text Mining-a feature of Web Intelligence to derive information from the unstructured textual data on the web and device the consensus based strategy to business decisions. Performance evaluations, which provide employers with an opportunity to assess their employees' contributions to the organization, are essential to developing a powerful work team. The primary goals of a performance evaluation system are to provide an equitable measurement of an employee's contribution to the workforce, produce accurate appraisal documentation to protect both the employee and employer, and obtain a high level of quality and quantity in the work produced.

6. Conclusion

We proposed the isoperimetric co-clustering Algorithm a new method for partitioning the document word bi par-title graph The Proposed algorithm requires a solution to a sparse system of linear equations. Experiments Performed demonstrate the advantage of our approach over spectral approach in terms of quality efficiency and stability in partitioning the Document-word bipartite graph. We present results on real data sets illustrating the effectiveness of our approach. The results show that the use of side-information can greatly enhance the quality of text clustering and classification, while maintaining a high level of efficiency.

7. Acknowledgement

We would like to sincerely thank Assistant Prof. **S.DhanaSekran** for his advice and guidance at the start of this article. His guidance has also been essential during some steps of this article and his quick invaluable insights have always been very helpful. His hard working and passion for research also has set an example that we would like to follow. We really appreciate his interest and enthusiasm during this article. Finally we thank the Editor in chief, the Associate Editor and anonymous Referees for their comments.

References

[1] C. C. Aggarwal and H. Wang," Managing and Mining Graph Data" New York, NY, USA: Springer, 2010.
[2] C.C.Aggarwal,"Social Network Data Analytics" NewYork, NY, USA: Springer, 2011.
[3] C. C. Aggarwal and C.-X.Zhai," Mining Text Data" NewYork, NY, USA: Springer, 2012
[4] C. C. Aggarwal and C.-X.Zhai, "A survey of text classification algorithms in Mining Text Data" New York, NY, USA: Springer, 2012
[5] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets," in Proc. SIAM Conf. Data Mining, 2005, pp. 358–369.
[6] Y. Zhou, H. Cheng, and J. X. Yu, "Graph clustering based on structural/attribute similarities," PVLDB, vol. 2, no. 1, pp. 718–729, 2009
[7] S. Zhong, "Efficient streaming text clustering," Neural Netw.vol. 18, no. 5–6, pp. 790–798, 2005
[8] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non- negative matrix factorization," in Proc. ACM SIGIR Conf. NewYork, NY, USA, 2003, pp. 267–273
[9] F. Sebastiani, "Machine learning for automated text categorization," ACM CSUR, vol. 34, no. 1, pp. 1–47, 2002.
[10] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," Inf. Syst., vol. 25, no. 5, pp. 345-366, 2000.
[11] H. Frigui and O. Nasraoui, "Simultaneous clustering and dynamic keyword weighting for text documents," in Survey of Text Mining, M. Berry, Ed. New York, NY, USA: Springer, 2004, pp. 45-70.
[12] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD, 2000, pp. 109-110.
[13] Y. Sun, J. Han, J. Gao, and Y. Yu, "iTopicModel: Information network integrated topic modeling," in Proc. ICDM Conf., Miami, FL, USA, 2009, pp. 493-502.
[14] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An efficient data clustering method for very large databases," in Proc. ACM SIGMOD Conf., New York, NY, USA, 1996, pp. 103-114.
[15] 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.

Author Profile

N.S.Krishna Prasad is currently a PG scholar in Computer Science and Engineering from the Department of Computer Science at Rajalakshmi Engineering College, Chennai. He received his Bachelor Degree in Computer Science and Engineering from Kalasalingam University, Krishnankoil, Srivilluputhur, Virdhunagar District and Tamilnadu. His Research areas include Data Mining and Computer Networks.

S.DhanaSekran is currently working as an Asst.Professor from the Department of Computer Science and Engineering at Rajalakshmi Engineering College, Chennai and Tamilnadu. He received his Bachelor Degree in Information Technology from Vel Tech Engineering College, Chennai and Tamilnadu. He received his Master Degree from SRM University, Chennai and Tamilnadu. His main research interests lie in the area of Software Engineering. He attended two international and two national conferences.