ISSN: 2321-7782 (Online)
e-ISJN: A4372-3114
Impact Factor: 7.327

Volume 7, Issue 3, March 2019

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Review Paper on Content and Side Information based Text Mining

Snehal S. Ingavale
Computer Department,
Government Residence Women Polytechnic,
Tasgaon, Sangali – India

Abstract: In various applications of text mining, side information is available along with each text documents. The side-information may be of different kinds, like the links in the document, user-access behavior from web logs, document provenance information or other non-textual attributes which are embedded into the text document. Such attributes may contain a large amount of information for clustering purposes. However, the relative importance of this side-information may be difficult to estimate, especially when some of the information is noisy. In such cases, it can be dangerous to incorporate side-information into the mining process, because it can either improve the quality of the representation for the mining process, or it can add noise to the process. Therefore way to perform the mining process, so as to maximize the advantages from using this side information. Existing system proposes classical partitioning algorithms with probabilistic models in order to create an effective clustering approach.

Key Words: Text Mining, Classification, Clustering, Side Information.

I. Introduction

Data Mining is known as the type of database analysis that attempts to extract useful patterns or relationships in a group of data. A major goal of data mining is to extract previously unknown useful relationships among different data.

A. Text Mining

Text mining, also known as text data mining, roughly equivalent to text analytics, 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 shows in Fig 1 of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), 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 (i.e., learning relations between named entities) [1].

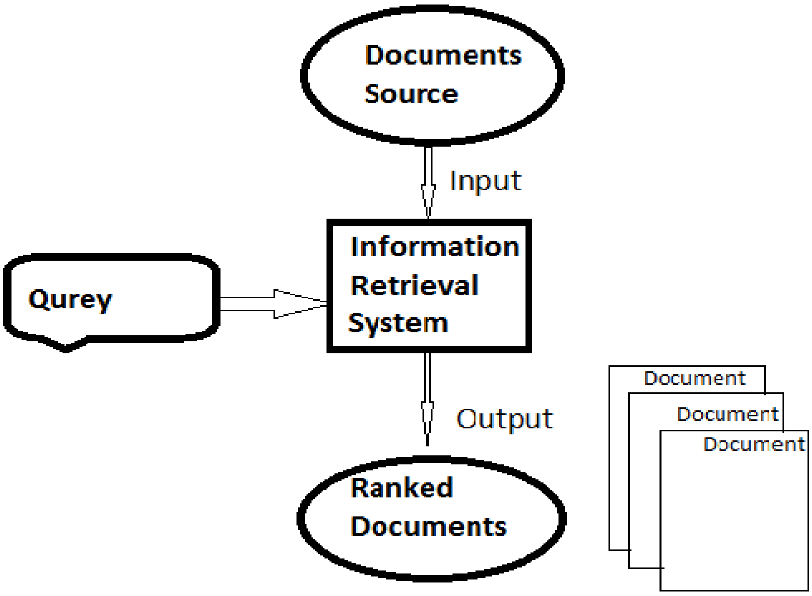

Fig.1Text Mining and Information Retrieval

B. *Side Information*

The problem of text mining arises in the context of many application domains such as the web, social networks, and other digital collections. A tremendous amount of work has been done in recent years on the problem of text collections in the database and information retrieval communities. However, this work is primarily designed for the problem of pure text collection, in the absence of other kinds of attributes. In many application domains, a large 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 mining process. Some examples of such side information are as follows.

1. *Meta-data*

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.

2. *Web logs*

In an application in which we track user access behaviour of web documents, the user-access behaviour may be captured in the form of web logs. For each document, the meta-information may correspond to the browsing behaviour 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.

3. *Links present in Text Document*

Text documents, 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.

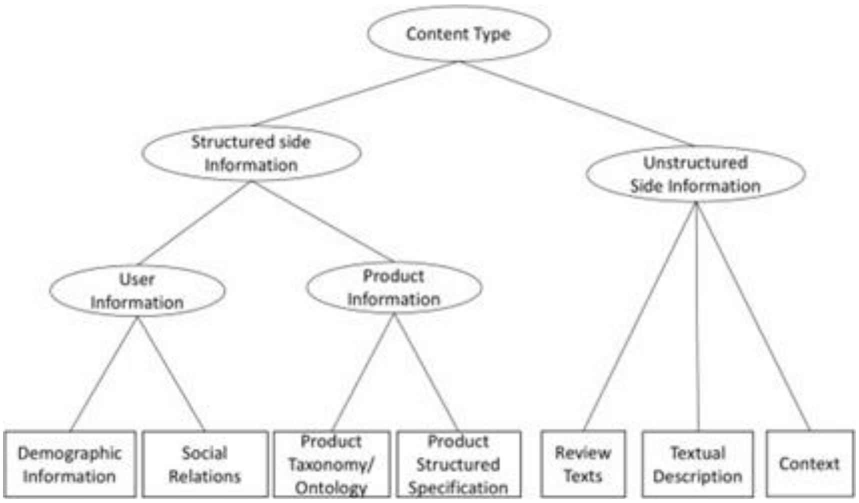

Fig.2 Classification of Side Information

II. Related Work

Charu C. Aggarwal, Yuchen Zhao and Philip S. Yu [1] designed an algorithm which combines classical Partitioning algorithms with probabilistic models in order to create an effective clustering approach. Then they show how to extend the approach to the classification problem. They presented experimental results on a number of real data sets in order to illustrate the advantages of using such an approach. They presented methods for mining text data with the use of side-information.

S. Guha, R. Rastogi, and K. Shim demonstrates [2] that for discovering groups and identifying interesting distributions in the underlying data clustering is used in data mining. Traditional clustering algorithms either favor clusters with spherical shapes and similar sizes. In this paper a clustering algorithm is presented which is called CURE that is more robust to outliers, and identifies clusters having non-spherical shapes and wide variances in size. A random sample drawn from the data set is first partitioned and each partition is partially clustered. The partial clusters are then clustered in a second pass to gain the desired clusters.

D.Cutting, D. Karger, J. Pedersen, and J. Tukey [3] explains the Hybrid Technique (Scatter-gather technique is the hybrid clustering technique). An example of the Scatter/Gather method, which provides a systematic browsing technique with the use of clustered document collection of the document organization. Initially the system scatters the collection of document into a small number of several document groups, or clusters, and presents short summaries of documents to the users. The user selects one or more of the groups for further study based on these summaries. The selected groups are gathered together to form a sub collection documents. The scatter-gather approach can be used for organized browsing of tremendous amount of document collections, because it creates a natural hierarchy of similar documents. However, these methods are designed for the pure text data clustering, and do not work for in which the text-data is combined with other forms of data.

T. Liu, S. Liu, Z. Chen, and W.Y. Ma[4] explains Feature extraction and feature selection techniques are used to reduce feature space dimensionality. In feature extraction it extracts a set of new features from original features through some functional mapping. In feature selection it chooses a subset from the original feature set according to some criteria. Document frequency, information gain, term strength are some of the feature selection methods. Unsupervised feature selection methods are much worse than supervised feature selection. In order to utilize the efficient supervised method an iterative feature selection method that iteratively performs clustering and feature selection is proposed in this paper.

S. Zhong demonstrates [5] that clustering data streams has been a new research topic, recently used in many real data mining applications, and has attracted a lot of research attention. However, there is not much work on clustering high-dimensional streaming text data. This paper merges an efficient online spherical k-means algorithm with an existing scalable clustering strategy to achieve fast and adaptive clustering of text streams. The OSKM algorithm modifies the spherical k-means algorithm, using online update based on the well-known.

III. Conclusion

In this paper, I presented methods for mining text data with the use of side-information. Many forms of text databases contain a large amount of side-information or Meta information, which may be used in order to improve the clustering process In order to design the clustering method, combination of an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information.

References

1. Aggarwal, C.C., Yuchen Zhao, Yu, P.S."On the Use of Side Information for Mining Text Data", Knowledge and Data Engineering, IEEE Transactions on, Vol. 26, No. 6, pp. 1415-1429, June 2014.
2. S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large Databases," in Proc. ACM SIGMOD Conf., New York, NY, USA, 1998, pp. 73–84.
3. 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.
4. 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.
5. S. Zhong, "Efficient streaming text clustering," Neural Netw., vol. 18, no. 5–6, pp. 790–798, 2005.
6. 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.

Author(s) Profile

Snehal S. Ingavale, received B.E.and M.E. Computer Engineering) from Savitribai Phule Pune University, India in 2013 and 2016 respectively. She is currently working as a Lecturer in Government Residence Women Polytechnic, Tasgaon, Sangli, India. Her research areas are data mining and web mining.