IJCAT - International Journal of Computing and Technology, Volume 4, Issue 3, March 2017
ISSN : 2348 - 6090
www.IJCAT.org

Relevant User Data Mining Using Auxiliary Information

[1] Shailesh Choudhari; [2] Dr. S. D. Choudhari

[1] M-Tech. CSE, Department, SBITM COE, Betul, India

[2] Principal SBITM COE, Betul, India

Abstract - In the digital era there is tremendous growth of digital information. This information also contains side information with it. The side information is the auxiliary information which may also be useful. The side information is of the type such as citation from the scientific publications, the authorship, the co-authorship, etc. such types of side information contain tremendous amount of information for performing clustering process. In this paper we deal with a principled way of performing mining process with the use of side information from the different documents. We make use of the clustering algorithm for the formation of clusters. After mining text data we search the keywords based on the user behavior.

Keywords: *Text mining, Side information, mining.*

1. Introduction

The rapid increase of digital information in the digital world creates problems in extracting relevant information from those huge data. The large amount of work on the text clustering has been done in recent years [18] [20]. In many text documents along with the text data, side information is also present. This side information is nothing but the auxiliary information which is useful for the clustering process. Some of the examples of the side information are as follows:-

a. Many text documents have links among them. These links are useful and contains a large amount of useful information for mining purposes. Such type of side information may provide the correlations among the documents which is not easily accessible from raw content.

b. Documents may be linked with user-tags in many network and user-sharing applications. This may also be quite informative.

We are making use of the side information for clustering the data for doing effective text mining. The process of deriving high quality information from text is known as text data mining. While this side-information can sometimes prove useful in improving the quality for the clustering process, but when the side-information is noisy it can be a risky approach and the quality of the mining process can actually become worse. Therefore, we will proceed towards to use an approach which carefully finds the well-organized form of the clustering characteristics of side information with that of text content. The basic approach of the system is that it can form a clustering in which the text attributes along with the side-information provide similar hints about the character of the basic clusters, and at the same time fail to consider those features in which conflicting hints are provided. Also we will use an efficient searching method based on user behavior. With the help of the use of this user behavior search we will get the more relevant searching results and the desired output.

2. Problem Statement

The problem statements consist of the drawbacks of the previous system and how those drawbacks are overcome in this research work.

- Current information retrieval systems do not provide results according to the user's individual needs and interests. When the clusters are formed we get huge amount of data. It is very difficult to find the information of the user's interest from these data. Our project allows the user to retrieve the files of his interests by providing the user behavioral searching.

- The auxiliary information present in the text documents also proves useful for the clustering process to give more refined clusters. The

COATES algorithm performs the clustering with the help of auxiliary information.

.

3. Implemented Methodology

Our objective will be firstly to collect information i.e. gathering the data set. After this extracting the keywords and the side information from those data sets will be our next objective. Once the keywords and the side information are extracted we will then perform the clustering using the COATES algorithm. Then the Text classification is carried out, means classifying the clustered text for generating the optimized result according to User Behavior (localization, personalization). Then the output will be shown in the form of graph. Graphical representation will show the relevant data mined from the particular page by removing the irrelevant information. Also the analytical mined reports will be generated. These reports will depend on the previous searching method and the method used in our proposed system.

3.1 Implementation Steps:

1. Gathering data set
2. Extracting keywords and side-information
3. Formation of cluster
4. Outcomes

Firstly we gathered large amount of data set from internet which includes the data on IPC Act of India and some generalized data. After this the preprocessing on data was done. Preprocessing is the very important thing since it may affect the result of clustering technique. Preprocessing includes the stop words removal techniques and the stemming technique. Once the preprocessing is done then the indexing of data has been started. Indexing is the process of maintaining the frequency of the particular keyword from the documents.

Stop Words: Stop words are the words which should be filtered out before or after the text processing. Stop words should be removed to support the phrase searching because they can cause problems while searching for phrases that include them. Some of the list of stop words includes, “is, the, at, which, on, are, but, onto, etc”. Such stop words are removed in the preprocessing step of our project.

Stemming: Stemming words are the derived words from their root words. Most of the times we come across words which have similar semantic interpretation and instead of those words their root words can be used for the information retrieval process. Stemming words includes, “Relating/Related/” can be replaced by their root word “Relate”. Such stemming words are replaced by their root words.

3.1.1 COATES Algorithm

The COATES algorithm is used for performing the clustering using the auxiliary information. COATES is the abbreviation of Content and Auxiliary attribute based Text clustering algorithm. The COATES algorithm takes k number of clusters as input. This algorithm can be applied only after preprocessing is performed i.e. the stop words are removed and the stemming is performed. COATES algorithm mainly works in two phases:

Initialization: Initialization is the first phase for performing the COATES algorithm. This phase uses a standard text clustering approach without the use of any side information. We use the K-means algorithm for this purpose. K-means algorithm is used because this algorithm is very simple and can provide quickly and effectively a starting point. In this first phase only pure text is used, and avoid auxiliary information.

Main Phase: The main phase is the phase after the execution of the initialization phase. This phase iteratively reconstructs the clusters by using the text and also the auxiliary information. The main phase uses the text content as well as the auxiliary information in order to improve the quality of the clustering process.
Steps of Algorithm Step 1: Normally clustering is performed without using any side information with the help of K means algorithm.

- The algorithm selects k random points as initial cluster centers.
- Each point from the given dataset is assigned to the closest cluster
- Each cluster center is then recomputed as the average of points in that cluster.
- Step ii and iii are repeated until the clusters are formed.

Step 2: Re-clustering is done with the help of auxiliary information.

This section gives the brief overview of the implementation details for the proposed system. In the proposed system the objectives are;
The first objective is gathering data set. From the large amount of data, the data on which we will be working is

selected. We choose some general data set for working and also some data on IPC Act of India. The second objective is extracting keywords and side information. Once the data set is gathered we then separated the keywords and the side information from the documents in the data set. The third objective is formation of clusters. After the extraction of keywords and the side information the clustering algorithm is applied. We apply the COATES algorithm. The details of the COATES Firstly we gathered large amount of data set from internet which includes the data on IPC Act of India and some generalized data. After this the preprocessing on data was done. Preprocessing is the very important thing since it may affect the result of clustering technique. Preprocessing includes the stop words removal techniques and the stemming technique. Once the preprocessing is done then the indexing of data has been started. Indexing is the process of maintaining the frequency of the particular keyword from the documents.

3.1.3 Recall and Precision Graph

The figure 3.1 shows the recall and the precision graph. The values are plotted in terms of the percentage. The blue color represents the existing system and the red color represents the developed system. The recall of the developed system is less than that of the existing system as our system shows the output on the basis of the personalized search. The formula for the calculation of the recall and the precision are as follows

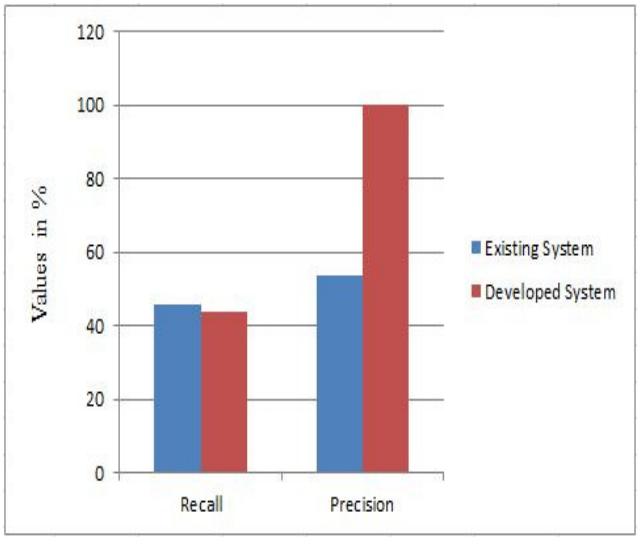

Figure 3.1: Recall and Precision Graph

$$\text{Recall} = \frac{A}{A+B}$$

$$\text{Precision} = \frac{A}{A+C}$$

Where A = the relevant record
B = the irrelevant retrieved record
C = the irrelevant record

3.1.4 Accuracy Graph of the System

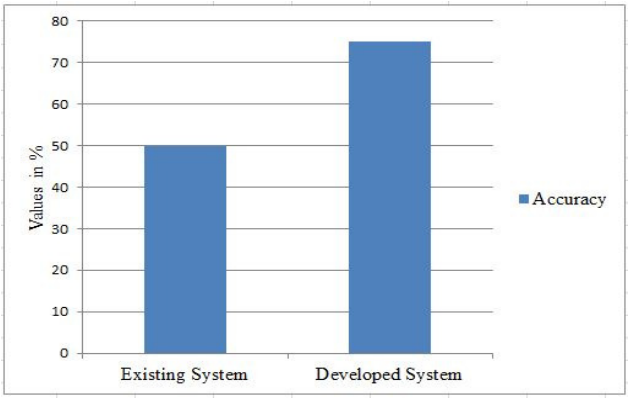

Figure 3.2: Accuracy Graph of the System

The above figure 5.9 represents the accuracy graph of the system. The accuracy of the existing system and the developed system are shown in terms of percentage. The accuracy of the system is calculated by taking the mean of the recall and the precision value. The formula is as follows,

Accuracy = (recall + precision) / 2

Here we shows that the techniques and the algorithms which we have used in our project works efficiently and gives more accurate results. We have compared two algorithms to find which one is more efficient and it shows that the COATES algorithm gives more accurate results. We have also compared our system with the existing system and found that the developed system gives more accurate results.

4. Future Scope

Here we specified how the system can be improved using other advance techniques.
The future work may include:

To identify the frequency patterns of the user by making the use of machine learning techniques like the artificial neural networks and many others. For example, given some existing data, we want to predict the behavior of the unseen data.

To make use of the association rules learning algorithms. These are the popular and well researched methods for discovering interesting relations between variables in the large databases.

5. Conclusion

There are many text mining domain applications which consist of auxiliary information along with the text documents. However, it may prove to be a difficult process for computing the importance of this auxiliary - information, especially when it is noisy and hence the quality of mining process may decrease. Therefore, we provide an efficient way to perform text mining which improve the text mining quality significantly. Here we focus on how to extract the side-information from the text document.

References

[1] Yuchen Zhao, Philip S and C. C. Aggarwal, "On the Use of Side Information for Mining Text Data", IEEE Transactions on knowledge and data engineering, vol. 26, issue 6, June 2014.

[2] Tarannum Bibi, Pratiksha Dixit Rutuja Ghule and Rohini Jadhav, "Web Search Personalization Using Machine Learning Techniques", IEEE International Advance Computing Conference (IACC), 2014.

[3] G. karypis and Ying Zhao, "Hierarchical Clustering Algorithms for Document Datasets", Data Mining and Knowledge Discovery, vol. 10, pp. 141-168, 2005.

[4] D. Napoleon and M.Praneesh, "An Efficient Numerical Methods for the Prediction of Clusters using K-means Algorithm with Bisection method for Comparing Uniform and Random Distribution Data Points", International Journal of Innovative Research in Computer and Communication Engineering, vol. 1, issue 8, October 2013.

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

[6] Y. Gong , W. Xu and X. Liu, "Document clustering based on nonnegative matrix factorization", Proceedings of ACM SIGIR Conference, New York, USA, pp. 267–273, 2003.

[7] S. C. Gates, P. S.Yu and C. C. Aggarwal, "On using partial supervision for text categorization", IEEE Transaction Knowledge and Data Engineering, vol. 16, issue 2, pp. 245–255, February 2004.

[8] P. S. Yu and C. C. Aggarwal, "A framework for clustering massive text and categorical data streams", Proceedings of SIAM Conference Data Mining, pp. 477–481, 2006.

[9] J. Zhang, Q. He, K. Chang and E. P. Lim, "Bursty feature representation for clustering text streams", Proceedings of SDM Conference, pp. 491–496, 2007.

[10] S. Zhong, "Efficient streaming text clustering", Neural Network., vol. 18, pp.790–798, 2005.

[11] G. P. C. Fung, J. X. Yu and H. Lu, "Classifying text streams in the presence of concept drifts", Proceedings of PAKDD Conference, Sydney, NSW, Australia, pp. 373–383, 2004.

[12] H. Wang and C. C. Aggarwal, Managing and Mining Graph Data. New York, USA: Springer, 2010.

[13] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections", Proceedings of CIKM Conference, New York, USA, pp. 778–779, 2006.

[14] J. Chang and D. Blei, "Relational topic models for document networks," Proceedings of AISTASIS, Clearwater, FL, USA, pp. 81–88, 2009.

[15] Q. Mei, D. Cai, D. Zhang and C.-X. Zhai, "Topic modeling with network regularization," Proceedings of WWW Conference, NewYork, USA, pp. 101–110, 2008.

[16] Y. Sun, J. Han, J. Gao and Y. Yu, "iTopicModel: Information network integrated topic modeling," Proceedings of ICDM Conference, USA, pp. 493–502, 2009.

[17] T. Yang, R. Jin, Y. Chi and S. Zhu, "Combining link and content for community detection: A discriminative approach," Proceedings of ACM KDD Conference, New York, USA, pp. 927–936, 2009.

[18] Y. Zhou, H. Cheng and J. X. Yu, "Graph clustering based on structural/attribute similarities," Proceedings of VLDB, vol. 2, issue 1, pp. 718–729,2009.

[19] Ting Yuan, Jian Cheng, Xi Zhang, ShuangQiu and Hanqing Lu, "Recommendation by Mining Multiple User Behaviors with Group Sparsity", Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[20] Ryen W. White, Wei Chu, Xiaodong He, Ahmed Hassan1, Yang Song and Hongning Wang, "Enhancing Personalized Search by Mining and Modeling Task Behavior", Proceedings of ACM International World Wide Web Conference (WWW), 2013.