ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 10, October 2015*

# A Review on Generation of Side Information Mining Text Data & Protection of Cyber Attack

**Rajesh A. Roy**[1], **Prof. P. S. Kulkarni**[2]

Department of Computer Science &Engineering, RCERT, Chandrapur[1,2]

**Abstract:** As the use of web side information for critical services has increased, the sophistication of attacks against these applications has grown as well. To protect web applications several intrusion detection systems have been proposed. In this paper, several techniques which are meant for detect ion of web application related attacks. The Intrusion detection system provides the following: .Monitoring and analyzing of user and system activity. Auditing of system configurations and vulnerabilities. Assessing the integrity of the files and critical system. Statistical analysis of activity patterns. Abnormal activity analysis. Operating system audit.

**Keywords:** Side information Intrusion detection operating system audit.

## I. INTRODUCTION

In various forms in online forums such as the web, social networks, and other information networks. In most cases, the data is not purely available in text form. A lot of side-information is available along with the text documents. Such side-information may be of different kinds, such as 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. However, the relative importance of this side-information may be difficult to estimate, for when some of the information is noisy. In such cases, it can be risky to incorporate side-information into the clustering process, because it can either improve the quality of the representation for clustering, or can add noise to the process. Therefore, we need a principled way to perform the clustering process, so as to maximize the advantages from using this side information. In this paper, we design an algorithm which combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach. We present experimental results on a number of real data sets in order to illustrate the advantages of using such an approach.

### 1.VISUAL SALIENCY WITH SIDE INFORMATION

We propose novel algorithms for organizing large image and video datasets using both the visual content and the associated side information, such as time, location, authorship, and so on. Earlier research have used side-information as pre-filter before visual analysis is performed, and we design a machine learning algorithm to model the join statistics of the content and the side information. Our algorithm, Diverse-Density Contextual Clustering (D2C2), starts by finding unique patterns for each sub-collection sharing the same side-info, e.g., scenes from winter. It then finds the common patterns that are shared among all subsets, e.g., persistent scenes across all seasons. These unique and common prototypes are found with Multiple Instance Learning and subsequent clustering steps. We evaluate D2C2 on two web photo collections from Flicker and one news video collection from TRECVID. Results show that not only the visual patterns found by D2C2 are intuitively salient across different seasons, locations and events, classifiers constructed from the unique and common patterns also outperform state-of-the-art bag-of-features classifiers.

We propose a D2C2 algorithm for visual pattern discovery by joint analysis of visual content and side information. A content collection is partitioned into subsets based on side information, and the unique and common visual patterns are discovered with multiple instance learning and clustering steps that analyzes across and within these subsets. Such patterns help to visualize the data content and generate vocabulary-based features for semantic classification. The proposed framework is rather general which can handle all types offside information, and incorporate different common/unique pattern extraction algorithms. One future work is to improve the generation of common patterns by emphasizing the shared consistencies, instead of the current heuristic clustering. Another future work is to explore other applications using the unique common patterns.

### 2.RECONSTRUCTING STORYLINE GRAPHS FOR IMAGE RECOMMENDATION FROM WEB COMMUNITY PHOTOS

In this paper, we investigate an approach for reconstructing storyline graphs from large-scale collections of Internet images, and optionally other side information such as friendship graphs. The storyline graphs can be an effective summary that visualizes various branching narrative structure of events or activities recurring across the input photosets of a topic class. In order to explore further the usefulness of the storyline graphs, we leverage them to perform the image sequential prediction tasks, from which photo recommendation applications can benefit. We formulate the storyline reconstruction problem as an inference of sparse time-varying directed graphs, and develop an optimization

DOI 10.17148/IJARCCE.2015.410108

algorithm that successfully addresses a number of key challenges of Web-scale problems, including global optimality, linear complexity, and easy parallelization. With experiments on more than 3.3 millions of images of 24 classes and user studies via Amazon Mechanical Turk, we show that the proposed algorithm improves other candidate methods for both storyline reconstruction and image prediction tasks. We proposed an approach for reconstructing storyline graphs from large sets of photo streams available on the Web. With experiments on more than three millions of Flicker images for 24 classes and user studies via AMT, we validated that our scalable algorithm can successfully create storyline graphs as an effective structural summary of large scale and ever-growing image collections. We also quantitatively showed the excellence of our storyline graphs for the two prediction tasks over other candidate methods.

Acknowledgement: This work is supported in part by NSFIIS-1115313, AFOSR FA9550010247, Google, and Alfred P. Sloan Foundation

## 3. Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots

Today internet has made the life of human dependent on it. Almost everything and anything can be searched on net. Web pages usually contain huge amount of information that may not interest the user, as it may not be the part of the main content of the web page. Web Usage Mining (WUM) is one of the main applications of data mining, artificial intelligence and so on to the web data and forecast the user's visiting behaviors and obtains their interests by investigating the samples. Since WUM directly involves in applications, such as, e-commerce, e-learning, Web analytics, information retrieval etc.

Weblog data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link and this information can be used in several applications like adaptive web sites, modified services, customer summary, pre-fetching, generate attractive web sites etc. There are varieties of problems related with the existing web usage mining approaches. Existing web usage mining algorithms suffer from difficulty of practical applicability. This paper continues the line of research on Web access log analysis is to analyze the patterns of web site usage and the features of users behavior. It is the fact that the normal Log data is very noisy and unclear and it is vital to preprocess the log data for efficient web usage mining process.

Preprocessing is the process comprises of three phases which includes data cleaning, user identification, and pattern discovery and pattern analysis. Log data is characteristically noisy and unclear, so preprocessing is an essential process for effective mining process. In this paper, a novel pre-processing technique is proposed by removing local and global noise and web robots. Preprocessing is an important step since the Web architecture is very complex in nature and 80% of the mining process is done at this phase. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web

Dataset are used for evaluating the proposed preprocessing technique. Web log data is a collection of huge information. Many interesting patterns available in the web log data. But it is very complicated to extract the interesting patterns without preprocessing phase. Preprocessing phase helps to clean the records and discover the interesting user patterns and session construction. But understanding user's interest and their relationship in navigation is more important. For this along with statistical analysis data mining techniques is to be applied in web log data. Data preprocessing treatment system for web usage mining has been analyzed and implemented for log data.

Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the failed HTTP status code and finally robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries. This preprocessing step is used to give a reliable input for data mining tasks. Accurate input can be found if the byte rate of each and every record is found. The data cleaning phase implemented in this paper will helps in determining.

## 4. Recommendation of Optimized Web Pages to Users Using Web Log Mining Techniques

Now a days, user rely on the web for information, but the currently available search engines often gives a long list of results, much of which are not always relevant to the user's requirement. Web Logs are important information repositories, which record user activities on the search results. The mining of these logs can improve the performance of search engines; since a user has a specific goal when searching for information. Optimized search may provide the results that accurately satisfy user's specific goal for the search.

In this paper, we propose a web recommendation approach which is based on learning from web logs and recommends user a list of pages which are relevant to him by comparing with user's historic pattern. Finally, search result list is optimized by re-ranking the result pages. The proposed system proves to be efficient as the pages desired by the user, are on the top in the result list and thus reducing the search time. An optimized recommendation system with two level architecture has been proposed in this paper. A matching query algorithm and Rank Updating algorithm have been proposed for implementing effective web search. It also contributed to the result improvement as the recommendation is based upon the users' feedback and analysis of web log.

The results show that the proposed system improves the relevancy of the pages and thus reduce the time user spends in seeking the required information. In future, we would like to develop an architecture which would provide the perfect relevancy of the query terms to the user.

DOI 10.17148/IJARCCE.2015.410108

## II. METHODOLOGY

We 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, we combined an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information. This general approach is used in order to design both clustering and classification algorithms. 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.

This algorithm as COATES throughout, which corresponds to the fact that it is a content and auxiliary attribute based Text clustering algorithm. We assume that an input to the algorithm is the number of clusters k. As in the case of all text-clustering algorithms, it is assumed that stop-words have been removed, and stemming has been performed in order to improve the discriminatory power of the attributes.

## III. CONCLUSION

We 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, we combined an iterative partitioning technique with a probability estimation process which computes the importance of different kinds of side-information. This general approach is used in order to design both clustering and classification algorithms. 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. 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.

## 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. New York, NY, USA: Springer, 2011

.[3] C. C. Aggarwal and C.-X. Zhai, Mining Text Data. New York, 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] C. C. Aggarwal and P. S. Yu, “A framework for clustering massive text and categorical data streams,” in Proc. SIAM Conf. Data Mining, 2006, pp. 477–481.

[6] C. C. Aggarwal, S. C. Gates, and P. S. Yu, “On using partial supervision for text categorization,” IEEE Trans. Knowl. Data Eng.,vol. 16, no. 2, pp. 245–255, Feb. 2004.

[7] C. C. Aggarwal and P. S. Yu, “On text clustering with side information,” in Proc. IEEE ICDE Conf., Washington, DC, USA, 2012.

[8] Chengyong Yang; Erliang Zeng; Tao Li; Narasimhan, G.,“ Clustering genes using gene expression and text literature data 2,” in Proc. IEEE 2005

[9] Amirul, M.; Omar, M.A.; Aini, N.; Karuppiah, E.K.; Mohanavelu; Soo Saw Meng; Poh Kit Chong, “Sorting very large text data in multi GPUs,” in Proc. IEEE 2012

[10] Suzuki, T.; Hayashi, K., “Text data compression ratio as a text attribute for a language-independent text art extraction method” ICDIM 2010.

DOI 10.17148/IJARCCE.2015.410108