# A Survey on Data Mining Techniques and Real Time Applications

Anu Mathews

**Abstract— With the proliferating amount of information in the World Wide Web, data mining techniques are gaining more importance. Recommender Systems have gained importance in this realm, as they facilitate the extraction of relevant information by providing the users with information which has been filtered based on the user's preferences. This paper describes the various data mining techniques and also different types of recommender systems prevalent today and also throws light on some of the real time applications of data mining and recommender systems.**

**Keywords— Contexts, data mining, prediction, recommender systems, social networks, web mining**

## 1 INTRODUCTION

WEB mining can be defined as the assimilation of information collected by conventional data mining techniques with information obtained over the World Wide Web. Web mining is a fast-growing research field. Web usage mining, Web structure mining, and Web content mining are the different aspects of Web Mining. Web usage mining refers to the detection of user access patterns by investigating Web usage logs. Discovering useful knowledge from the structure of hyperlinks is the focus of Web structure mining. Web content mining aims to extract useful information from web page contents. The major focus of Web Content Mining consists of the following:

**Information retrieval**: Information retrieval deals with the extraction of structured data from Web pages, such as data referring to products and search results. Obtaining useful data allows a service provider to provide services. Common methods used for information extraction include machine learning and Natural Language processing based techniques.

**Web information integration and schema matching**: Similar information will be represented differently by different websites. One of the very significant problems with many efficient applications is the identification of semantically identical data.

**Opinion extraction from online repositories:** Quite a lot of online opinion sources like customer reviews of products, blogs, forums, and chat rooms are available these days. Marketing efficiency and product quality gauging rely on extracting of opinions.

**Knowledge synthesis**: Generation of concept hierarchies manually takes tremendous amount of time. Synthesizing and arranging the information on the Web to give the user a consistent picture of the topic domain can be achieved by exploiting methods that probe the information redundancy of the Web.

**Segmenting Web pages and detecting noise**: A user generally would not be interested in advertisements, navigation links, etc., but rather would be interested in the main content of the Web page. Obtaining the relevant content of the pages by automatically segmenting Web page is an interesting problem.

## 2 DATA MINING TECHNIQUES

Data mining processes involve one or more of these techniques:

**Pattern Recognition**: Recognizing patterns in data sets is an interesting problem. Recognizing some deviation in your data happening at regular intervals, or a repetitive pattern of a certain variable over time are common aspects of pattern recognition. For example, you might see that the sales of a certain product seem to spike just before the school reopening time or notice that a change to a particular weather drives more people to your website.

**Classification**: Data mining techniques which gather various attributes together into distinguishable categories, which can be used to draw conclusions, or meet some target fall into the category of Classification. For example, suppose we are evaluating data on individual customers' economic backgrounds and purchase histories, in which case, each user may be classified as "low," "medium," or "high" credit risks, which can throw light on the user's profile

**Association:** Association is more specific to dependently associated variables. In this case, events or attributes that are highly correlated with another event or attribute, is the major focus . For example, a user who buys a specific item, often tends to buy a second, related item (eg: bread and butter, computer and printer). This is usually what's used for item recommendation in online stores.

**Detection of outliers**: Recognizing the pattern may not give a clear understanding of the data set. Identifying anomalies or outliers in the data is also very important. For example, a data measurement which was continuously giving a reading of 10, immediately gives a reading of 100. The user might want to analyze the spike and see potential causes behind it, so as to

• *Author Anu Mathews is currently working as an Assistant Professor in Department of Computer Science & Engineering, T. John Institute of Technology, Bengaluru, India*

have a better understanding of the concerned system being studied.

**Clustering:** Clustering comprises of grouping pieces of data together based on their similarities. For example, based on how much expendable income users have, or how frequently they tend to shop at a particular store, a data analyst can choose to cluster different statistics of a user into different packets.

**Regression:** Regression, is fundamentally a form of modeling which is used to spot the probability of a certain variable, given the presence of other variables. For example, based on other factors like availability, consumer demand, and contention, regression can be used to estimate a certain price. Discovering the exact relationship between two or more variables in a particular data set is the major focus of regression.

**Prediction:** Prediction is yet another important data mining technique, because it is used to give an estimate about the types of data that might be seen in the future. A mere insight into historical trends is sufficient to come up with a somewhat accurate prediction of what might happen in the future. For example, credit histories and past purchases of customers can be used to predict whether they will be a credit risk in the future.

## 3 Recommender Systems

In our day to day life we need many things to be searched over the internet. Search engines are available for this purpose. Every time we search something, we try to get the most relevant results, and this can be achieved using Recommender systems. In a world where there is a wide variety of choices, recommender systems help users find and evaluate items of interest. Recommender systems work by connecting users with items to "consume" by relating the content of recommended items or the opinions of other individuals with the consuming user's actions or opinions.

The two basic approaches used by Recommender systems in general are collaborative filtering and content-based filtering. Collaborative filtering arrives at a recommendation that's based on a model of previous user behavior. This type of filtering works by constructing a model from a single user's behavior or also from the behavior of other users who have similar traits. When other users' behavior is considered, collaborative filtering utilizes group knowledge to form a recommendation based on like users. In essence, recommendations are based on an automatic collaboration of multiple users and filtered on those who exhibit similar preferences or behaviors. Content-based filtering makes an item recommendation by first comparing the user's profile with the contents of that item. For example, suppose a user is frequently visiting a site and purchasing some products. Then the characteristics of those items can be used to identify and recommend similar items to this user.

## 4 State Of The Art

It's a long-standing challenge to develop efficient Web mining systems. In this section, some of the real time applications which have deployed data mining techniques are being mentioned.

In "A Collaborative Decentralized Approach to Web Search" by Athanasios Papagelis and Christos Zaroliagis, bottom-up approach was adopted to study the web dynamics based on users' feedback [1]. The procedures and mechanism of a peer-to-peer, bottom-up search engine that can provide search results by combining the users' preference for web pages has been explored here. Personalized results were also incorporated with inclusion of necessary extensions. An alternative version of PageRank that combines the link analysis and the users' preference is obtained by integrating this approach with PageRank.

An algorithm to obtain social groups, which outperforms the solutions manually configured by users has been proposed in "A Comprehensive Study on Willingness Maximization for Social Activity Planning with Quality Guarantee", by Hong-Han Shuai, De-Nian Yang, Philip S. Yu, and Ming-Syan Chen [2]. Due to complications of social connectivity and the diversity of possible interests among friends, the process of coordinating social group activities manually is tedious and time-consuming for users. To address these issues, automatic selection and recommendation of potential participants of a social group activity has been done. A new problem, named Willingness mAximization for Social grOup (WASO) has been identified. However, the solution obtained by a greedy algorithm is likely to be captured in a local optimal solution.

Content and context-specific information are prevalent in Social media networks. Content and context information are rich sources of information for mining, and the full power of mining and processing algorithms can be realized only with the use of a combination of the two. In "Exploring Context and Content Links in Social Media: A Latent Space Method", Guo-Jun Qi, Charu Aggarwal, Qi Tian, Heng Ji and Thomas S. Huang propose a novel algorithm which mines both context and content aspects in social media networks to discover the hidden semantic space [3]. The use of any multimedia extraction algorithms can be facilitated with the help of this mapping of the multimedia objects into latent feature vectors. By mining the geometric structure concealed in the content links between multimedia objects, this algorithm effectively solves the problem of scanty context links when compared to the latest methods in analysis of multimedia. By concurrently utilizing both context and content information based on hidden structure between correlated semantic concepts, we can directly build annotation models for multimedia annotation.

The huge acceptance of the Social Web has converted social networks to very good means of advertising mediums. But, all the potential of these mediums has not been exploited by many of the current social publicity strategies, because they obstruct users' online life: the social contexts in which they are involved. In "Inferring Contexts from Facebook Interactions: A Social Publicity Scenario" by Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Ariasupta , a model to infer users' social contexts by the application of several Natural Language Processing (NLP) and data mining techniques over users' interaction data on Facebook is used to reverse this situation [4]. An exposure to users' social life allows ads to target the most potential customers, which in

turn is beneficial for both companies and potential customers.

In "Flexible Frameworks for Actionable Knowledge Discovery", by Longbing Cao, Yanchang Zhao, Huaifeng Zhang, Dan Luo, Chengqi Zhang and E.K. Park, a formal view of actionable knowledge discovery (AKD) from the system and decision-making perspectives has been proposed [5]. AKD is drafted to give operable business rules that can be persistently linked with business processes. It is a closed optimization problem solving process from problem statement, design of framework to actionable pattern discovery. To aid such processes, four types of generic AKD frameworks have been put forward: Postanalysis-based AKD, Unified-Interestingness-based AKD, Combined-Mining-based AKD, and Multisource Combined-Mining-based AKD (MSCM-AKD). A real-life case study of MSCM-based AKD is validated to extract debt prevention patterns from social security data. Many sophisticated problems and applications can be dealt with the proposed models by extracting actionable deliverables for immediate decision making.

Crowdsourcing techniques have been used by multimedia and social computing research to annotate objects, actions and scenes in social video sites like YouTube. But crowdsourcing of personal and social traits in online social video or social media content has not been addressed predominantly. In "Mining Crowdsourced First Impressions in Online Social Video" Joan-Isaac Biel and Daniel Gatica-Perez address the problems of crowdsourcing the annotation of first responses of video bloggers (vloggers) personal and social qualities in conversational YouTube videos, and (2) modeling the collaboration of different vlogger aspects by mining the responses [6].

The reputation of social networking has spread from the Internet to mobile domains. The collaborative work of internet with cellular networks and self-organized mobile ad hoc networks gives way to advanced pervasive social networking (PSN) at any time and in any location. For protecting crucial social activities and aiding reliable social computing and data mining, security of data communications is vital in PSN. Trust has an important role in PSN for complementary activities among strangers. It aids people in overthrowing notions of uncertainty and risk and engages in trusted social behaviors. In "Protect Pervasive Social Networking Based on Two-Dimensional Trust Levels" by Zheng Yan and Mingjun Wangwe utilize two dimensions of trust levels gauged by either a reliable server or individual PSN nodes or both to control PSN data access in a heterogeneous manner based on attribute-based encryption [7]. Under relevant system and security models, this proposed scheme is highly efficient and also secure.

In "Uploader Intent for Online Video: Typology, Inference, and Applications", Christoph Kofler, Subhabrata Bhattacharya, Martha Larson, Tao Chen, Alan Hanjalic, and Shih-Fu Chang investigate the reason for which a user has uploaded a particular video on the Internet [8]. Users upload video for particular reasons. But the reasons behind a particular upload will be seldom stated explicitly in the video metadata. Information about the reasons inspiring uploaders has the potential eventually to benefit a wide range of application areas, including video production, video-based advertising, and video search. Here, a combination of social-Web mining and crowdsourcing has been applied to arrive at a typology that characterizes the uploader intent of a broad range of videos. Multimodal features, including visual semantic features, are indicative of uploader intent. This can be utilized in order to classify videos automatically into uploader intent classes.

## 5 Conclusion

In this paper many of the recent web mining techniques have been discussed. The importance of Recommender systems is also exposed with relevant examples.

## References

[1] JAthanasios Papagelis and Christos Zaroliagis, "A Collaborative Decentralized Approach to Web Search", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 42, NO. 5, SEPTEMBER 2012

[2] Hong-Han Shuai, De-Nian Yang, Philip S. Yu, and Ming-Syan Chen, "A Comprehensive Study on Willingness Maximization for Social Activity Planning with Quality Guarantee", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 1, JANUARY 2016.

[3] Guo-Jun Qi, Charu Aggarwal, Qi Tian, Heng Ji and Thomas S. Huang"Exploring Context and Content Links in Social Media: A Latent Space Method", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 34, NO. 5, MAY 2012

[4] Sandra Servia-Rodríguez, Ana Fernández-Vilas, Rebeca P. Díaz-Redondo, and José J. Pazos-Ariasupta, "Inferring Contexts From Facebook Interactions: A Social Publicity Scenario", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 6, OCTOBER 2013

[5] Longbing Cao, Yanchang Zhao, Huaifeng Zhang, Dan Luo, Chengqi Zhang and E.K. Park, "Flexible Frameworks for Actionable Knowledge Discovery", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 9, SEPTEMBER 2010

[6] Joan-Isaac Biel and Daniel Gatica-Perez, "Mining Crowdsourced First Impressions in Online Social Video", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 16, NO. 7, NOVEMBER 2014

[7] Zheng Yan and Mingjun Wangwe, "Protect Pervasive Social Networking Based on Two-Dimensional Trust Levels", IEEE SYSTEMS JOURNAL, VOL. 11, NO. 1, MARCH 2017

[8] Christoph Kofler, Subhabrata Bhattacharya, Martha Larson, Tao Chen, Alan Hanjalic, and Shih-Fu Chang, "Uploader Intent for Online Video: Typology, Inference, and Applications ", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 17, NO. 8, AUGUST 2015.