National Workshop-Cum-Conference on Recent Trends in Mathematics and Computing (RTMC) 2011
Proceedings published in International Journal of Computer Applications® (IJCA)

Indirect Correlations: Significance in Web Mining

Indu Singh [1], Gagandeep kaur[2]

Astt. Professor[1, 2]
Department of Computer Science & Engineering
JCDM,college of Engg.
Sirsa (Haryana) – India

ABSTRACT

“Direct” association rules reflect relationships existing between items that relatively often co-occur in common transactions direct association rules are dedicated to describe the direct correlations among the items in a frequent item set, indirect association rules are dedicated to describe the indirect correlations between the two items in a infrequent item set. Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. When a pair of items, (A, B), which seldom. occur together in the same transaction, are highly dependent on the presence of another item set C, then pair (A, B) are said to be indirectly associated via C .In this paper indirect association rules and significance of web usage mining are explained. How association’s rules are beneficial for web usage mining is explained.

1. INTRODUCTION

Association rules mining is one of the most important and widespread data mining techniques. They reflect regularities in the co-occurrence of the same items within a set of transactions. A classical example of the association rule is the discovery of sets of products usually purchased together by many independent buyers. In the web environment,
association rules are typically applied to HTTP server log data that contain historical user sessions. Web sessions are gathered without any user involvement and, additionally, they reliably reflect user behavior while navigating throughout a web site. For that reason, web sessions can be regarded
as an important source of information about users.
Association rules that reveal similarities between web pages derived from user behavior can be simply utilized in recommender systems. The main
goal of such a recommendation is to suggest to the current user some web pages that appear to be useful.[1]

2. WEB MINING

Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data). There is a purpose to adding the extra clause about structure and usage data. The reason being that mining Web content by itself is no different than general

data mining, since it makes no difference whether the content was obtained from the Web, a database, a file system or through any other means. Web content can be variegated, containing text and hypertext, image, audio, video, records, etc. Mining each of these media types is by itself a sub-field of data mining. The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined[2].

2.1 Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Text mining and its application to Web content has been the most widely researched. Some of the research issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. [2]

2.2Web structure mining

Web structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. Web Structure Mining can be regarded as the process of discovering structure information from the Web. [2]

2.3 Web usage mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered.[2]

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. As mentioned before, the mined data in this category are the secondary data on the Web as the result of interactions. These data could range very widely but generally we could classify them into the usage data that reside in the Web clients, proxy servers and servers [13] The Web usage mining process could be classified into two commonly used approaches [14]. The first approach maps the usage data of the Web server into relational tables before an adapted data mining technique is performed. The second approach uses the log data directly by utilizing special pre-processing techniques. As is true for typical data mining applications, the issues of data quality and pre-processing are also very important here. The typical problem is distinguishing among unique users, server sessions, episodes, etc. in the presence of caching and proxy servers [15; 13]. For the details and comparison of some pre-processing methods for Web usage data we refer to [31]. In general, typical data mining methods (see for example in [13]) could be used to mine the usage data after the data have been pre-processed to the desired form. However, modifications of the typical data mining methods are also used such as composite association rules [16]. Web users would be interested in, among others, techniques that could learn their information needs and preferences, which is user modeling possibly combined with Web content mining. On the other hand, information providers would be interested in, among others, techniques that could improve the effectiveness of the information on their Web sites by adapting the Web site design or by biasing the user's behavior towards satisfying the goals of the site. In other words, they are interested in learning user navigation patterns. Then the learned knowledge could be used for applications such as personalization (at a Web site level), system improvement, site modification, business intelligence, and usage characterization (see [13] for the detail). [2]

2.4 Main Techniques of Web Usage Mining

2.4.1 Association

To determine what are the set of page views often accessed together in the same server

2.4.2 Clustering

To find groups of users who share similar browsing behavior.

2.4.3Classification

To categorize Web users according to their past access history.

3. ASSOCIATION RULE MINING

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. It makes correlation among items that are grouped into transactions, deducing rules that define relationships between item sets. The rules have a user-stipulated support, confidence, and length. Association rule mining has attracted tremendous attention from data mining researchers and as a result several algorithms have been proposed for it . [1]

Let I = {i1, i2, …., im} be the collection of all the items and D be the set of database transactions where each transaction T is a set of items such that T €I. Let A be a set of items. A transaction T is said to contain A if and only if A T. An association rule is an implication of the form A -> B, where A €I, B €I, and A ∩ B = φ. They are two terms associated with association rules.
These are: Support and Confidence.
If the support of itemset {AB} is 30%, it means "30% of all the transactions contain both the itemsets – itemset A and itemset B".
Support of item set{AB} = Count Of the transactions containing the itemsets A and B
Total Number of Transactions If the confidence of the rule is 70%, it means "70% of all the transactions that contain itemset A also contain itemset B".
Confidence of the rule A -> B = Support A/Support AB.[1]

Association rules are further classified in two categories, which are as follows:

3.1 Direct Association Rules

These shows attribute value conditions that occur frequently together in a given dataset.
A direct association rule is the relationship X→Y, where X € D, Y €D and X ∩ Y = ∅ [3].
For direct Association the algorithms can be used are Apriori algorithm, FP Growth method, Incremental Algorithm

3.2indirect Association Rules

These show the indirect relationship between two items, which are associated with third item which is called transitive item. They are further classified in two types:

3.2.*1 Partial indirect association rules*

A partial indirect association rule is described by partial indirect confidence conP#(di→P#dj, dk) as follows:

conP#(di→P#dj,dk)=con(di→dk)·con(dk→dj) (1) [3].

3.2.2 Complete indirect association rules

The complete indirect association rule di→#dj aggregates all partial indirect association rules from di to dj with respect to all existing transitive pages dk € Tij and is characterized by complete indirect confidence con#(di→#dj).A complete indirect association rule from di to dj exists if and only if there exists at least one partial indirect association rule from di to dj , i.e., Tij ≠∅[3]

4. APPLICATION OF WEB USAGE MINING

The Web and identifies links that are potentially interesting to the user. The Web Watcher starts with a short description of a user's interest. Each page request is routed through the Web Watcher proxy server in order to easily track the user session across multiple Web sites and mark any interesting links. Web Watcher learns based on the particular user's browsing plus the browsing of other users with similar interests.

4.1 Personalization

Personalizing the Web experience for a user is the holy grail of many Web-based applications e.g. individualized marketing for e-commerce. Making dynamic recommendations to a Web user, based on her/his profile in addition to usage behavior is very attractive to many applications, e.g. cross-sales and up-sales in e-commerce. Web usage mining is an excellent approach for achieving this goal. Web server logs were used by Yan et. al. [6] to discover clusters of users having similar access patterns. The system proposed in [6] consists of an offline module that will perform cluster analysis and an online module which is responsible for dynamic link generation of Web pages. Every site user will be assigned to a single cluster based on their current traversal pattern. The links that are presented to a given user axe dynamically

Selected based on what pages other users assigned to the same cluster have visited. The Site Helper project learns user’s preferences by looking at the page accesses for each user. A list of keywords from pages that a user has spent a significant amount of time viewing is compiled and presented to the user. Based on feedback about the keyword List, recommendations for other pages within the site are made.

4.2 System Improvement

Performance and other service quality attributes axe crucial to user satisfaction from services such as databases, networks, etc. Similar qualities are expected from the users of Web services. Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission load balancing, or data distribution. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins, etc. Web pages requested from a particular user or a group of users accessing from the same proxy server. The locality measure can then be used for deciding pre-fetching and caching strategies for the proxy server. The increasing use of dynamic content has reduced the benefits of caching at both the client and server level. These profiles are then used to pre-generate dynamic HTML pages based on the current user profile in order to reduce latency due to page generation [13]

4.3Site Modification

The attractiveness of a Web site, in terms of both content and structure, is crucial to many applications, e.g. a product catalog for e-commerce. Web usage mining provides detailed feedback on user behavior, providing the Web site designer information on which to base redesign decisions. While the results of any of the projects could lead to redesigning the structure and content of a site, the adaptive Web site project (SCML algorithm) [7; 8] focuses on automatically changing the structure of a site based on usage patterns discovered from server logs. Clustering of pages is used to determine which pages should be directly linked[13].

4.4 Business Intelligence

Information on how customers axe using a Web site is critical information for marketers of e-tailing businesses. Buchner et al have presented a knowledge discovery process in order to discover marketing intelligence from Web data. They define a Web log data hypercube that will consolidate Web usage data along with marketing data for e-commerce applications[13].

4.5 Usage Characterization

While most projects that work on characterizing the usage, content, and structure of the Web don't necessarily consider themselves to be engaged in data mining, there is a large amount of overlap between Web characterization research and Web Usage mining. Catledge et al.[9] discuss the results of a study conducted at the Georgia Institute of Technology, in which the Web browser Xmosaic was modified to log client side activity. The results collected provide detailed information about the user's interaction with the browser interface as well as the navigational strategy used go browse a particular site. The project also provides detailed statistics about occurrence of the various client side events such as the clicking the back/forward buttons, saving a file, adding to bookmarks etc. Pitkow et al. [10] propose a model which can be used to predict the probability distribution for various pages a user might visit on a given site. This model works by assigning a value to all the pages on a site based on various attributes of that page.[13]

5. WEB USAGE MINING AND ASSOCIATION RULES

Indirect Association provides an alternative approach to capture interesting infrequent patterns, For Web data, indirect association represents the distinct interests of Web users who share similar traversal path, Such patterns cannot be easily found using standard association and clustering techniques. Indirect Association can be used to group together patterns into more compact structures, Navigation pages form the mediators. Indirect association also provides a methodology to group together the discovered patterns according to the items they have in common. Even though the idea of grouping association patterns is not new, our work differs from others in terms of the types of patterns being grouped and how the combined patterns are represented[5].

6. CONCLUSION

With the information overload, Web mining is a new and promising research issue to help users in gaining insight into overwhelming information on the Web. Workshops on Web mining have been already or will be held to discuss its principle, architecture and algorithm in several international conferences As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it., Indirect Association provides an alternative approach to capture interesting infrequent patterns For Web data, indirect association represents the distinct interests of Web users who share similar traversal path, Such patterns cannot be easily found using standard association and clustering techniques .Indirect Association can be used to group together patterns into more compact structures. Navigation pages form the mediators. This paper explained a review of web usage mining and indirect associations in web data.

7. REFERENCE

[1] Weimin Ouyang and Qinhua Huang, "*Discovery Algorithm for Mining both Direct and Indirect Weighted Association Rules*" In International Conference on Artificial Intelligence and Computational Intelligence, 2009, IEEE.

[2] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining : Concepts and Applications & Rerearch Directions",in April,2009.

[3]Przemyslaw Kazienko,*"Mining Indirect Association Rules For Web Recommendation" Int. J. Appl. Math. Compt. Sci., Vol. 19, No. 1, 165–186.2009.*

[5]Pang-Ning Tan and Vipin Kumar," *Mining Indirect Associations in Web Data",* in the proceeding *of* conference on WEBKDD in 2001, LNAI 2356, pp. 145–166, 2002.

[6] T. Yah, M. Jacobsen, H. Garcia-Molina, and U. Dayal. From user access patterns to dynamic hypertext linking. In *Fifth International World Wide Web Conference,* Paris, France, 1996.

[7] Mike Perkowitz and Oren Etzioni. Adaptive web sites: Automatically synthesizing web pages. In *Fifteenth National Conference on Artificial Intelligence,* Madison, WI, 1998.

[8] Mike Perkowitz and Oren Etzioni. Adaptive web sites: Conceptual cluster mining. In *Sixteenth International Joint Conference on Artificial Intelligence,* Stockholm, Sweden, 1999.

[9] L. Catledgeand J. Pitkow. Characterizing browsing behaviors on the world wide web. *Computer Networksand ISDN Systems,* 27(6), 1995.

[10] Chi E. H., Pitkow J., Mackinlay J., Pirolli P., Gossweiler, and Card S. K. Visualizing *the evolution of web ecologie*s. In *CHI '98,* Los Angeles, California, 1998.

[11] R. Kosala and H. Blockeel, "Web Mining Research: A Survey,"*ACM SIGKDD* Explorations, vol. 2, no. 1, pp. 1–15. *2000.*

[12] M. Spiliopoulou. Data mining for the web. In *Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '99,* pages 588-589, 1999.

[13] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations,* 1(2), 2000.

[14]Borges and M. Levene. Data mining of user navigation patterns. In *Proceedings of the WBBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, San Diego, CA, USA,* pages 31-36, 1999.

[15] B. Masand and M. Spiliopoulou. Webkdd-99: Workshop on web usage analysis and user profiling. *SIGKDD Explorations,* 1(2), 2000.

[16] J. Borges and M. Levene. "*Mining association rules in hypertext databases".* In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), August 27-31, 1998, New York City, New York, USA,* 1998.