International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

Investigation of E-Commerce based User's Traversal Path Patterns using MFB and MBP

[1]E.Sandeep Krupakar & [2]Dr. A.Govardhan
[1]Research Scholar, Rayalaseema University, Kurnool.
Registration No: PP COMP.SCI. &ENG.0379
[2]Professor & Principal, Department of Computer Science & Engineering
JNTUH, Hyderabad

Abstract: *Web mining can be connected to discover user traversal path patterns that can be broke down to enhance traversability and structure of a website. It uncovers the data that how the web pages are become to and what the web clients are looking for. The web log server stores users' behavioural navigational pattern and it is the hotspot for traversal design inventions. In this paper we investigate the e-commerce based Traversal Path Patterns and purchasing behaviour using data mining technique, here we analyses two types of traversal patterns i.e Maximal Forward Path and Maximal Backward Path. A Maximal forward path is defined as the longest consecutive sequence of forward references before the first backward reference is made to visit some previously visited page in the same session. Where documents are interconnected together to enable collaborating access in e-commerce website design which is prime and significant concern for successful business certainly users traceability and accessibility made easy.*

Keywords: Web Mining, Traversal Path Patterns, Web log Server, Maximal Forward Path (MFP) and Maximal Backward Path (MBP).

1. Introduction

In web mining investigating user's traversal pattern path finding and users purchasing behaviour is essential using data mining technique, where documents are interconnected together to enable collaborating access in e-commerce website design is prime and significant concern for successful business. Certainly users can traceability and accessibility of traversal behaviour is made using mining techniques. The easy to move around on a website is important task. If the case if customer will not reach to targeted object in time or quickly what they looking for, it leads to web usability may cause to high web traffic which effects on sales and decrease the revenue drastically. Main challenging task is here to identify the user's traversal pattern for website its prime concern due to accessibility and performance which depends on the site structure. In addition to website structure and linkage, positioning, formatting and naming of links can greatly affect the navigation of users.

When users looking for fascinated item over e-commerce website, then they may travel from the one web page to another web page using hyperlinks provided. The ultimate goal is to improve the business revenue and decrease website usability by

making an effective website design by capturing and mining the user's traversal patterns. Where traversal paths are useful to obtain frequent patterns thus traversal path patterns can reveal valuable information for improving web site structure and navigability of web site.

2. Related Work

Web data mining has been a well-liked area of research in recent years. Different research areas in web mining have been considered. One of the hot topics of the research is web data mining in e-commerce.

Nivedita Roy et al. [3] describe integration of web mining to knowledge discovery process, its potential applications and techniques. Also, they present integrated architecture showing contribution of web mining to e-business by new technologies allowing better business decisions to be taken and some commercially available architecture is presented.

M. Spiliopoulou et al. [4] propose a methodology to improve the success of web sites based on the exploitation of navigation pattern discovery. The success of a web site affects and reflects directly the success of the company in the electronic market. They describe that success is modeled on the basis of the navigation behavior of the site's users. Also they exploit a navigation pattern discovery miner to study how the success of a site is reflected in the users' behavior and suggest how the site should be improved.

S. Ansari et al. [1] describe architecture for integrating e-commerce and data mining that can dramatically reduce the efforts for pre-processing, cleaning and data understanding in knowledge discovery. They emphasize on requirement for data collection at application server layer in order to support logging of data and metadata essential to discovery process.

Aditi Todi et al. [5] present approach for classification of data to benefit consumers and companies both in e-commerce. Two algorithms are useful for classification: Naive Bayes and Decision Tree. They show that Decision tree performs better than Naïve Bayes. Using this, competitors can understand how their competitors are priced and consumers understand what qualities of products are available.

Yang Hao et al. [6] describe the building process of warehouse platform on ecommerce system including system design, data preprocessing and OLAP analysis.They present implementation of data warehouse within e-commerce system and focuses on real time data warehouse modeling on web platform.

Weiyang Xu et al. [2] focus on algorithm research on web log in web usage mining perspective and get result sets for user frequent Access Path. They use log records from web server for analysis, clean it and then generate user session for data mining. Using mining algorithm, all user frequent access paths are found which support personalized services in e-commerce.

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 14
November 2017

3. System Study

3.1 User Traversal Pattern Mining

In web data mining, process initiated from web server log information which is captured from the web user access information. For extracting an effective pattern web log data must be per-processed towards cleaning unnecessary, duplicate and null data from the web log data set. While preprocessing towards remove the sparsity we use machine learning algorithms i.e Naive Bayes algorithm which process and give the best result for unstructured web log data. Later session has been extracted towards user identity and session identity for user traversal pattern mining we used web spider it's a program, that can collect and analyses all the web pages from the website using hyperlinks, more over all the analysed pages are classified as either content page or index page, where index pages indicates several hyperlinks to traversal on other webpages remaining category called content page consist of users traversal navigation information often visited and most liked

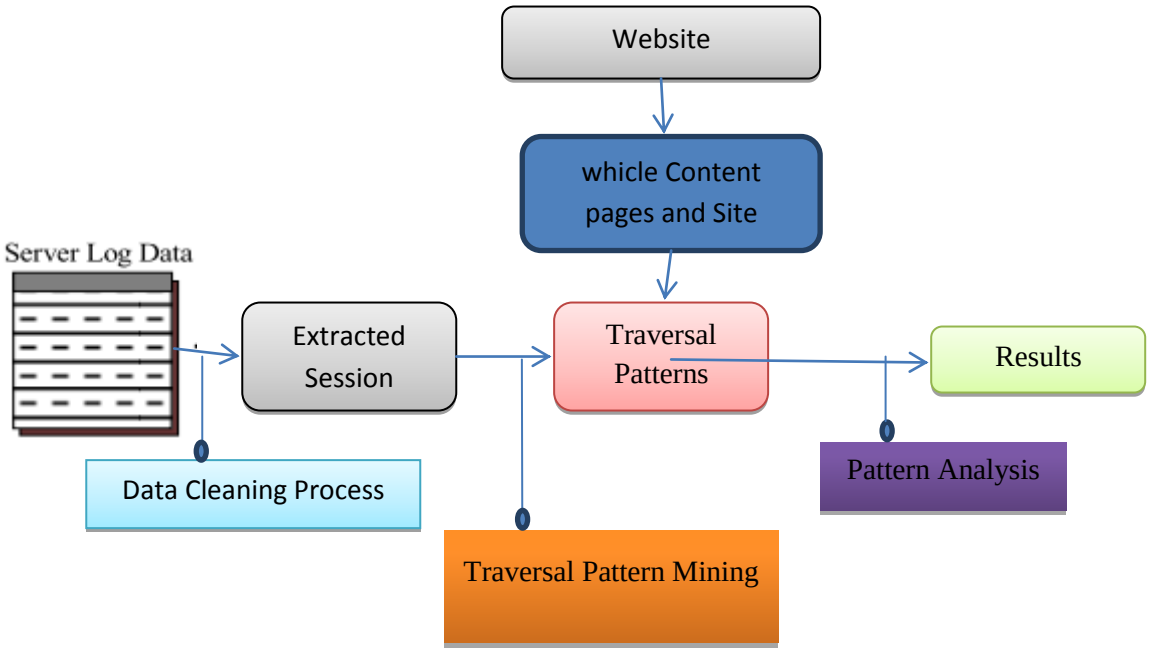


Figure 1: User Traversal Pattern Mining

Access patterns are appropriate to evolvement the web site structure design. Usually, visitors access set of pages consecutively that is logged in web server log. The web server log is preprocessed to clean and remove information that is irrelevant to obtain traversal patterns. Subsequently preprocessing level, user sessions are extracted. A user session is a collection of pages visited by the specific user in listed time. Since the extracted sessions, traversal patterns are establish and analyzed. Though classifying the web pages, the site structure is also deliberated as it signifies navigation structure. Figure 1. Describes the user traversal path pattern mining.

3.2 User Traversal Behavior Analysis using Web Log Mining

Web Usage Mining implies to find regular as well as incredible customer accessibility patterns from web browsing information that are kept in web server log. Web usage mining concentrates on policies that might forecast individual actions while the customer connects to the internet. Inside web log evaluation the primary rate of interests has been the customer as well as session recognition and also series of web pages being accessed by customers.

When an individual browses the web pages, request transfer to the web server whenever the client accesses new web pages or sources [7]. On the distant possibility that the online firms could find the following patterns of the visitors, they could anticipate customers' see patterns and also target market on an event of individuals.

When a customer browses the internet site web pages, demands are sent out to the web server whenever the individual accesses new web pages or sources. Web usage mining concentrates on strategies that might forecast customer behaviors while the customer connects to the internet. Inside web log evaluation the main passions have been individual as well as session recognition as well as series of web pages being accessed by customers.

4. Analysis

4.1 Traversal Pattern Discovery: MFP and MBP

In traversal pattern discovery we have been identified two kinds of traversal patterns named Maximal Forward Path (MFP) and Maximal Backward Path (MBP). Major composition web has constructed with natural graph, where pages are linked via hyperlinks. Although surfing the web, The visitor may perhaps travel forward along the graph by choosing a hyperlink in the existent page. He may likewise move backwards to any page went to already in a similar session click by back button. A forward reference shows the coveted data client is searching for.

MFP characterizes the traversal information more difficult to focus on duplicate of back tracking activities. Online browsing development isn't a straightforward single directional action but it may be a dual-directional activity. In arrange to channel out the redundant pattern from the log source, Maximal Forward Reference may be an idea of a maximal forward moving moment on web archives. All the in reverse traversal activities as it were happen to clients within the prepare of looking for Web pages that truly intrigued them. Hence, it is expected that as it were the forward browsing movement or Forward Reference is reflecting users' true browsing designs. When navigating web site, backtracking happens in case it is fundamental. It is much favored to fair click a connect to urge back to where you were prior, rather than hitting the browser back button a billion times. Finding client backtracking behaviors is one of the applications of traversal designs and can be utilized to make strides the site navigability.

Where MBP exhibits gatherings of nodes in the backward structure navigation. It shows a decent sign of how well the establishment of a site is developed and organized. The more extended the mix of nodes MBP holds the less

sorted out a site seems, by all accounts, to be. This can be translated as clients experiencing issues in finding their desired nodes and consequently they are compelled to peruse each connection in a steady progression, keeping in mind the end goal to limit the potential outcomes. On the probability that the MBP contains various similar navigation then this can advise that this specific reference of linkages is to a great extent made.

In a maximal forward or backward path, a succession of weblog sections begins with a similar client in the subsequent request and no passage seems more than once. The MFP contains passages produced from navigating forward to new unvisited pages, and the MBP includes sections created from navigating backwards to pages that have been gone by previously. The forward and backward references show up in the substitute request in the log. A client begins with a forward reference; they may have a back reference, they may have another forward text, and so on. The beginning of two maximal forward references might be indistinguishable, as a client may explore forward five pages, at that point backward two pages, at that point forward again three pages

4.2 MFP (Maximum Forward Path)

Here a given a collection of nodes are organized in hierarchal order that move from top order to bottom down. When the first back movement takes place the forward movement is dismissed. This results in a collection of nodes which is noticeable as maximum forward path.

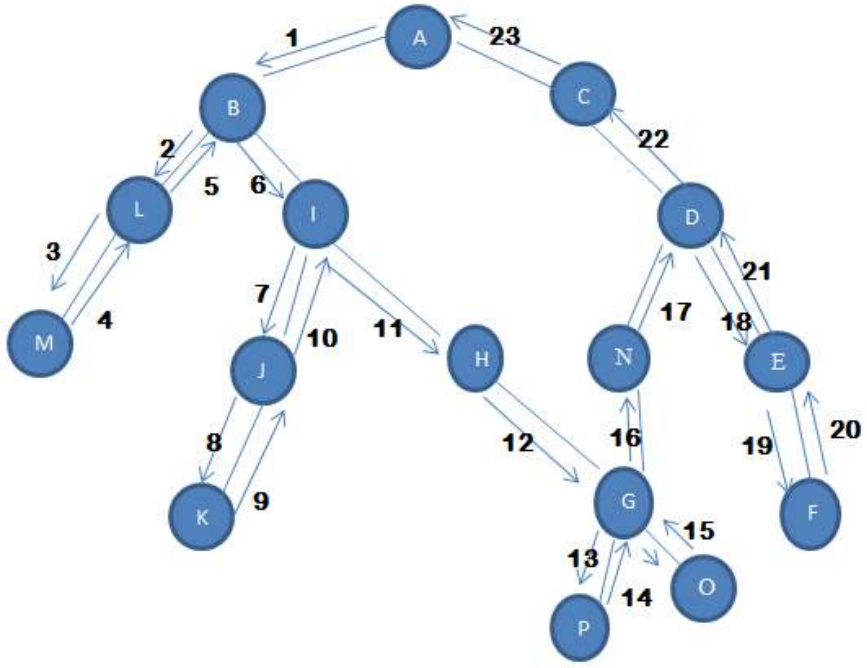
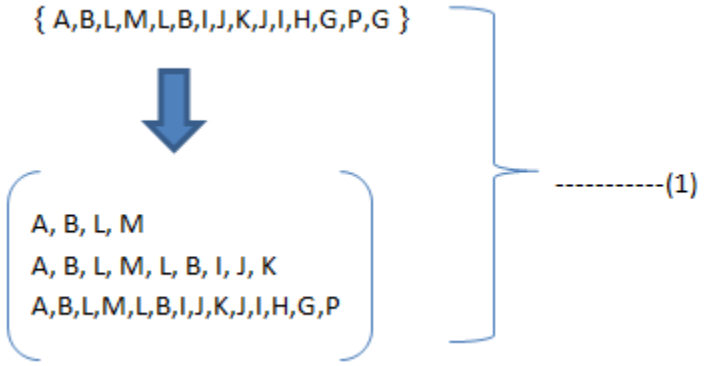


Figure 2 Traversal Patterns for the W W.W

E.g., from the above illustration we consider node from root node “A” to bottom down at node “G”

Node Traversal list= {A,B,L,M,L,B,I,J,K,J,I,H,G,P,G }

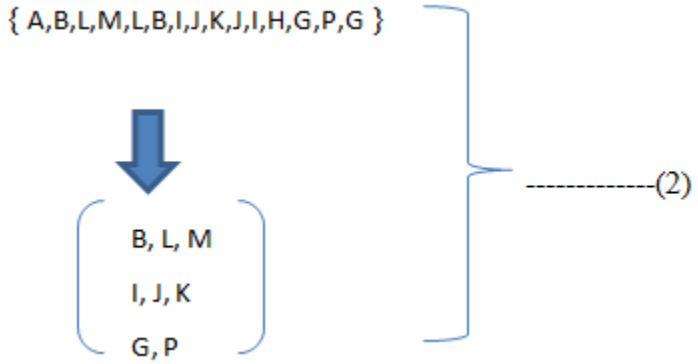
As per definition presented above, the maximum forward path for this instance will be taken out as below:



Here it demonstrates that the MFP is deep-rooted on the nodes M, K and P where back word movement starts taking place. Later travelling from node "A" to "G" yields 3 extreme forward paths enumerated above. Meanwhile MFP ignores all the backing directional travelling; it will cover only the nodes taken throughout the forward visits.

4.3 MBP (Minimum Backward Path)

Here a given a collection of nodes are organized in hierarchal order that throughout a specific session in time, the MBP begins at a node when a Backward traversal path occurs and move back to the node where a new forward movement was raised .Minimum Backward Path is "not" essential the back order of a maximum forward path. Over using as an example, the MBP for travelling from node "A" to "G", is scheduled as follows:



After comparing the traversal path among MFB and MBP i.e. (1) and (2), MBP nodes contains bidirectional movement, where MFP consist with own directional nodes

Algorithm to find maximal forward references

Here we develop two essential algorithms called Progressive Comprehensive Scan and Enriched Discriminating Scan for mining traversal patterns

Finding Maximal Forward References

For finding maximal forward reference in e-commerce traversal web log database which consist of hyperlinks traversed and referrer log in.Following are the traverse pattern set { (s_1,d_1), (s_2,d_2), (s_3,d_3), (s_4,d_4), (s_5,d_5).. (Sn,d_n) Of user, and be about to map in to the multiple subsequences, each of which signifies a maximal forward reference. The algorithm for finding all maximal forward reference is specified as follows. The traversal log details sorted by user id's, resulting in traversal path " { (s_1,d_1), (s_2,d_2), (s_3,d_3), (s_4,d_4), (s_5,d_5).. (Sn,d_n)}", for each user, where pairs of (s_i,d_i) are ordered by time, Algorithm MF is then applied to each user path to determine all of its maximal forward references. Let D_F Denote the database to score all resulting maximal forward references obtained.

Algorithm to find maximal forward references

Step 1: Set i=1 and String Y to null for initialization, where string Y is used to store the current forward reference path .Also, set the flag F=1 to indicate a forward traversal .

Step 2: Let A=s_i and B=d_i

if A is equal to null then

/* this is the beginning of a new traversal */

Begin

Write out the current string Y(if not null) to the database D_f

Set string Y=B;

Go to step 5;

End

Step 3: if B is equal to some reference (say j-th reference to) in string Y then

Begin

If F is equal to 1 then write out string Y to database D_F;

Discard all the reference after the j-th one is string Y;

F=0;

Go to step 5.

end

Step 4: Otherwise, append B to the end of string Y.

/* we are continuing a forward Traversal */

If F is equal to 0, set F=1.

Step 5: Set i = i+1; if the sequence is not completed scanned then go step 2.

Consider the traversal scenario in Fig. 1 for example. It can be verified that the first backward reference is encountered in the M to B At that point, the maximal forward reference ABLM is written to D_F (by Step 3). In the next move (i.e., from L to B), although the first conditional statement in Step 3 is again true, nothing is written to D_F since the flag F = 0, meaning that it is in a reverse traversal. The subsequent forward references will put ABIJK into

the string Y, which is then written to D_F when a reverse reference (from K to J) is encountered. The execution scenario by algorithm MF for the input in Fig. 3 is given in

Table 1. An Example Execution by Algorithm MFP

Move	String Y	Output D_F
1	AB	-
2	ABL	-
3	ABLM	-
4	ABL	ABLM
5	AB	-
6	ABI	-
7	ABI	-
8	ABIJ	-
9	ABIJK	-
10	ABIJ	ABIJK
11	ABIH	-
12	ABIHG	-
13	ABIHGP	-
14	ABIHG	ABIHGP
15	ABIHG	-
16	ABIHGO	-
17	ABIHG	ABIHGO

4.4 Traversal Pattern and Purchasing Behavior in E-Commerce

In e-commerce, traversal path patterns represent the navigation behavior of customers. The information about purchasing behavior of customer can be used to find association between purchasing items and this can help in improvement of cross selling. Considering both traversal pattern and purchasing behavior of customer can add value to association rule finding. Figure 6.4 demonstrates traversal and purchase behavior of customer. The customer traverses and purchases item in following order. First customer starts from A and goes to B where purchases item1. Then, customer sequentially visits L, M On G, customer purchases item2.

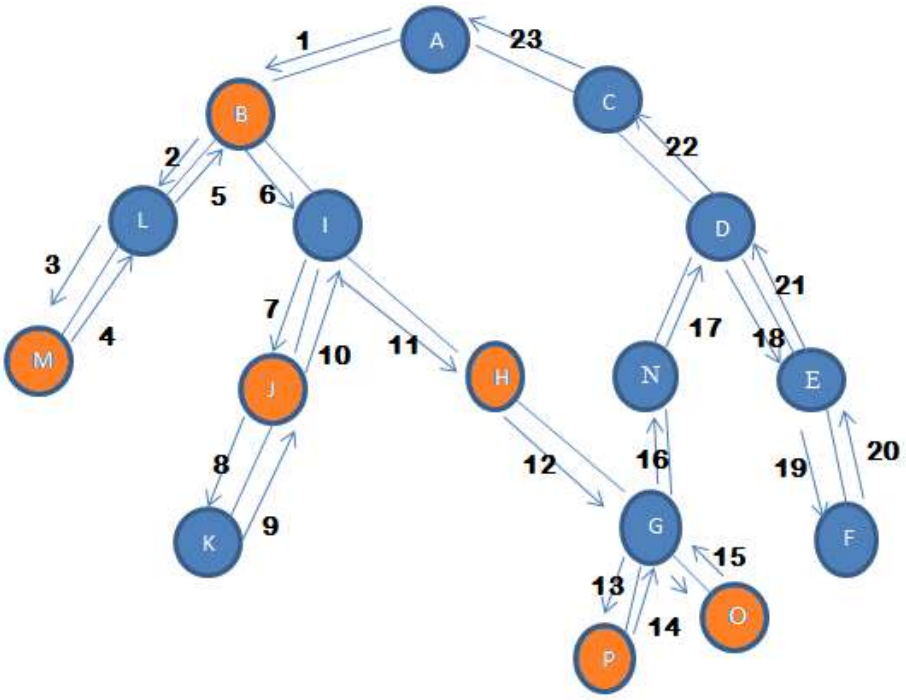


Figure 3: Demo on Traversal and Purchase Behavior of Customer

In another sequence, Customer visits A, then visits B where customer purchases item1. Next, from B, customer goes to I and then to J where purchases item3. Then, visited K Back to J, customer purchases item 4. Same way, other traversal sequences are followed. The customer transaction detail with respect to this traversal and purchase behavior is represented in Table 2

Table 2. Traversal and purchase behavior

Traversal Path	Purchase Details
A-B-L-M	B(item1) ,M(Item 2)
A-B-I-J-K	B(Item 1),J(Item 3)
A-B-I-H-G-P	B(Item 1),H(Item 4),P(Item 5)
A-B-I-H-G-O	B(Item 1),H(Item 4),O(Item 6)

Considering both traversal path pattern and purchase behavior is important for finding association between purchasing items. Following statement indicates that considering traversal path A-B-I-J-K , customer purchases item1 at B then generally also purchases item 3 at J

ABIJK [B(item1)=>Q(item3)].

Support for this can be defined as p/n where n is total number of customers and p indicates number of customers who traverse ABIJK and purchases item1 on B and then purchases item3 on J. Now support of ABIJK [B(item1)] is

q/n where n is total number of customers and q indicates number of customers who traverses ABIJK and purchase item1 on B. The confidence of ABIJK [B(item1)=>Q(item3)] can be found as under.

Confidence of ABIJK [B(item1)=>Q(item3)]

=Support of ABIJK [B(item1), Q(item3)] / support of ABIJK [B(item1)]

=p/q

The confidence describes that if customer follows traversal path ABIJK and purchases item1 on B then the probability of purchasing item3 on webpage Q is p/q.The support for traversal path pattern indicates number of customers following that path while traversing. If support for traversal pattern matches with minimum support defined by user then it is called frequent traversal pattern.

5. Conclusion

Web mining can be applied to find user traversal path patterns that can be analyzed to improve navigability and structure of web site. It reveals the information that how the web pages are accessed and what the web users are seeking for. The web server log stores browsing behavior of site visitors and it is the source for traversal pattern finding. The web server log is preprocessed to clean and eliminate information that is not required to find traversal patterns. Cleaned Web log is analyzed to identify visit sessions, each of which constitutes a basic processing unit for the discovery of interesting, prominent access patterns. A user session is a set of pages visited by the same user within the duration of one particular visit to a web site. Each user session consists of only the pages visited by a user in a row. We detect two types of traversal patterns the Maximal Forward Path and Maximal Backward Path.

A maximal forward path is defined as the longest consecutive sequence of forward references before the first backward reference is made to visit some previously visited page in the same session. Thus, the last reference in a maximal forward sequence indicates a content page that is desired by the user. The Maximal Backward Path demonstrates groups of nodes in the backward sequence combination. This presents a good indication of how well the infrastructure of a site is constructed and arranged.

6. References

[1] S. Ansari, R. Kohavi, L. Mason, Z. Zheng, "Integrating e-commerce and data mining: architecture and challenges", Proceedings IEEE International Conference on Data Mining, pages 27 – 34, 2001.

[2] Weiyang Xu, Zhenji Zhang, Shidong Zhou, "Research on mining algorithm of frequent access path of e-commerce users", IEEE International Conference on Service Operations and Logistics, and Informatics, ISBN : 978-1-4244-2013-1, Vol 1, pages 182-185, 2008.

[3] Nivedita Roy, Tapas Mahapaatra, "Web mining: a key enabler in e-business", Proceedings of International Conference on Services Systems and Services Management, ISBN: 0-7803-8971-9, Vol 2, pages 1121 – 1125, 2005.

[4] M. Spiliopoulou, C. Pohle, "Data Mining for Measuring and Improving the Success of Web Sites", Journal Data Mining and Knowledge Discovery, Volume 5,Issue 1-2, pages 85-114, 2001.

[5] Aditi Todi, Anahita Agrawal, Ankit Taparia, Nikhlesh Lakhmani, Dr. Rajashree Shettar, "Classification of E-Commerce Data Using Data Mining", International Journal of Engineering Science & Advanced Technology, Vol 2, Issue 3, pages 550–554, 2012.

[6] Yang Hao, Song Hongwei, Zhang Zili, "The application of e-commerce system based on data warehouse", Information Technology and Artificial Intelligence Conference (ITAIC), ISBN: 978-1-4244-8622-9, Vol 2, pages 493 – 496, 2011.

[7] Fang X. and Liu Sheng O.R. "LinkSelector: A Web Mining Approach to Hyperlink Selection for Web Portals," ACM Transactions on Internet Technology,Vol 4, Issue 2, pages 209-237, 2004.