WEB CONTENT MINING: A CASE STUDY FOR BPUT RESULTS

Binayak Panda[1], K Murali Gopal[2], Sudhanshu Shekhar Bisoyi[3]

[1]*Assistant Professor, Information Technology, GIET, Gunupur, Odisha, India*
[2]*Associate Professor, Computer Science and Engineering, GIET, Gunupur, Odisha, India*
[3]*Assistant Professor, Computer Science and Engineering, GIET, Gunupur, Odisha, India*

Abstract

In today's competitive world paper less work is gaining utmost importance. For this to happen role of Web Based Systems are incomparable. Different sectors like banking, retail are fully ported towards Web Based Systems, whereas the education sector is also not far behind to them. All most all universities or institutions are providing their own web portal for notification of news related to seminar/workshop/examination/result. In this article we have considered web portal of BPUT, Odisha with url http://results.bput.ac.in. More specifically we put our interest on the way results are being published or displayed. In this web portal for some cases the results are being displayed in an unorganized manner over multiple pages. On this unorganized data we are applying the concepts of Web Content Mining and providing a Web Content Mining tool for an organized access/view to the above said web contents.

Keywords: *Web Based System, Web Content, Web Content Mining, Web Content Mining Tool, Organized view of unorganized Web Content*

----------***----------

1. INTRODUCTION

[1]Web mining is a concept inherited for Data mining. Data mining is a tool that can extract predictive information from large quantities of data, and is data driven. Data mining is a process of inferring knowledge from a huge data set. The World Wide Web is also consists huge amount of unstructured or unorganized data. Web Mining is a process of collecting and integrating data from multiple websites.

[2]TOPICS OF WEB MINING
It is often said that the Web offers an unprecedented opportunity and challenge for data mining. We believe that this is so due to the following characteristics of the Web:

1. The amount of data/information on the Web is huge and still growing rapidly. Web data is also easily accessible.
2. The coverage of Web information is wide and diverse. One can find information about almost anything on the Web.

3. Data of all types exist on the Web, e.g., structured tables, texts, multimedia data (e.g., images and movies), etc.

4. Information on the Web is heterogeneous. Multiple Web pages may present the same or similar information using completely different formats or syntaxes, which makes integration of information a challenging task.

5. Much of the Web information is semi-structured due to the nested structure of HTML code and the need of Web page designers to present information in a simple and regular fashion to facilitate human viewing and browsing.

6. Much of the Web information is linked. There are links among pages within a site, and across different sites. These links serve as an information organization tool and also as indications of trust/authority in the linked pages and sites.

7. Much of the Web information is redundant. The same piece of information or its variations may appear in many pages or sites. This property has been explored in many Web data mining tasks.

8. Above all, the Web is a virtual society. It is not only about data, information and services, but also about interactions among people, organizations and automatic systems.

9. The Web is dynamic. Information on the Web changes constantly. Keeping up with the changes and monitoring the changes are important issues for many applications.

We can see why the Web is such a fascinating place and why it offers so many opportunities for web data mining.

[3] Web mining can be defined as mining of the World Wide Web (WWW) to find useful knowledge about user behavior, content, and structure of the web. It involves application of data mining techniques on the contents of WWW but is not limited to it.

From the Figure 1 classification of Web Mining as follows:
Web Structure Mining: is the technique to analyze and explain the links between different web pages and web sites. It mainly focuses on developing web crawlers. It works on hyperlinks and mines the topology of their arrangement.

Web Content Mining: focuses on extracting knowldge from the contents or their descriptions. It involves techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior.

Web Usage Mining: It focuses on digging the usage of web contents from the logs maintained on web servers, cookies logs, application server logs etc. It works on how and when user moves from one type of content to other. Thus, it can provide association between different contents.

[4] Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. More precisely , Web Con-tent Mining is that part of Web Mining which focuses on the raw information available in web pages; source data mainly consist of textual data in web pages (e.g., words, but also tags);

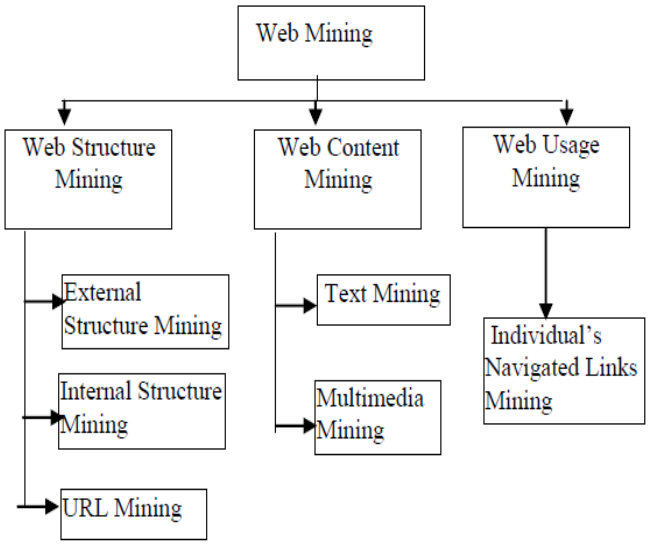


Fig-1: Classification of Web Mining

[5]Web Content Mining is the process of extracting useful information from the contents of Web documents. It may consist of text, images, audio, video information which is used to convey to the users about that documents. 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. Web content mining issues in term of Information Retrieval (IR) and Database (DB) view verses data representation, method and application categories is discuss and summarized in . While extracting the knowledge from images - in the fields of image processing and computer vision - the application of these techniques to Web content mining has not been very rapid.

2. PROPOSED WORK

In this work we have focused on the Text Mining, a sub class of Web Content Mining to be applied on http://results.bput.ac.in/ and a tool for organized display of result from the unorganized display.

2.1 Problem Statement

A student took admission in the year 2010 will appear his 3rd semester examination as highlighted in the table-1. This is the case if a student fails to pass in the first attempt. We have not included the Special examination which is being conducted once in a year. If that will be included then number of attempts can be made by a student will be more.

Table-1: [Result Display Link]

Semester Result Link (A 2010 admitted student can appear for exam of 3rd semester Subjects as specified.)	2010 Batch	2011 Batch	2012 Batch	2013 Batch
	1st Sem.			
	2nd Sem.			
	3rd Sem.	1st Sem.		
	4th Sem.	2nd Sem.		
	5th Sem.	3rd Sem.	1st Sem.	
	6th Sem.	4th Sem.	2nd Sem.	
	7th Sem.	5th Sem.	3rd Sem.	1st Sem.
	8th Sem.	6th Sem.	4th Sem.	2nd Sem.
		7th Sem.	5th Sem.	3rd Sem.
		8th Sem.	6th Sem.	4th Sem.
			7th Sem.	5th Sem.
			8th Sem.	6th Sem.
				7th Sem.
				8th Sem.

2.2 Mathematical Explanation of Problem

If a student of 2010 batch got 2 backlogs in his 3rd semester examination and he appears for the same with consequent batch students, where he clears 1 backlog with his juniors and the other with his sub juniors. Then with the current web portal his result will be spread over three different web pages.

Mathematically the above information can be said as,

1. R_1= {$RESULT_1$ from the result of 3rd semester link of 2010 batch},
2. R_2= {$RESULT_2$ from the result of 3rd semester link of 2011 batch},
3. R_3= {$RESULT_3$ from the result of 3rd semester link of 2012 batch}.

To get the Final/Updated result, the following operation should be carried out

$$R=F (R_1 \cup R_2 \cup R_3).$$

Where F is the function implemented as "BPUT RESULT CONTENT MINING TOOL".

Algorithm:

Step 1: Let SR_i be the Subject Result of i^{th} subject of a particular semester. i.e. SRi Ɛ Ri

Step 2: For each SR_i in the subjects of a particular semester Do

Consider the updated result from R1, R2 and R3 put in R.

Done

2.3 Explanation through Case Study

Let us consider a particular student of 2010 batch the results are spread over different pages like below

From Figure-2 we can see the Page with original Result

BIJU PATNAIK UNIVERSITY OF TECHNOLOGY, ORISSA

Result for B.TECH 1st Semester Examination, 2010-2011

Student Roll No: 100xxxx208
Student Name: xxxxxx xxxx xxxx
College: xxx xxx xxx
Branch: xxx xxx xxx
Published on: 24-Jun-2011

Sl.No.	Subject Code	Subject	Credit	Grade	Date
1	BE2101	BASIC ELECTRONICS	3	F	24-06-11
2	BE2104	MECHANICS	3	F	24-06-11
3	BE2105	PROGRAMMING IN C	3	F	24-06-11
4	BS1101	MATHEMATICS - I	4	F	24-06-11
5	BS1102	PHYSICS - I	3	D	24-06-11
6	HM3101	COMMUNICATIVE ENGLISH	2	F	24-06-11
7	BE7102	WORKSHOP PRACTICE	2	E	24-06-11
8	BE7103	PHYSICS LABORATORY	2	A	24-06-11
9	BE7105	BASIC ELECTRONICS LABORATORY	2	A	24-06-11
10	BE7107	PROGRAMMING IN C LABORATORY	2	A	24-06-11
11	HM7101	COMMUNICATIVE ENGLISH LABORATORY	2	A	24-06-11
		Total Credits:	28	**SGPA:** 4.54	

Fig-2: [Initial result]

Form Figure-3 we can see the Different web pages for the result of subject BE

LINK	GRADE
http://results.bput.ac.in/108_RES/100xxxx208.html	F
http://results.bput.ac.in/149_RES/100xxxx208.html	F
http://results.bput.ac.in/195_RES/100xxxx208.html	S
http://results.bput.ac.in/204_RES/100xxxx208.html	F
http://results.bput.ac.in/253_RES/100xxxx208.html	C

Fig-3

From Figure-4 we can see the Different web pages for the result of subject MECH

LINK	GRADE
http://results.bput.ac.in/108_RES/100xxxx208.html	F
http://results.bput.ac.in/149_RES/100xxxx208.html	F
http://results.bput.ac.in/204_RES/100xxxx208.html	F
http://results.bput.ac.in/253_RES/100xxxx208.html	F
http://results.bput.ac.in/322_RES/100xxxx208.html	D

Fig-4

From Figure-5 we can see the Different web pages for the result of subject CP

LINK	GRADE
http://results.bput.ac.in/108_RES/100xxxx208.html	F
http://results.bput.ac.in/195_RES/100xxxx208.html	F
http://results.bput.ac.in/253_RES/100xxxx208.html	C

Fig-5

From Figure-6 we can see the Different web pages for the result of subject MATH1

LINK	GRADE
http://results.bput.ac.in/108_RES/100xxxx208.html	F
http://results.bput.ac.in/195_RES/100xxxx208.html	F
http://results.bput.ac.in/253_RES/100xxxx208.html	F

Fig-6

From Figure-7 we can see the Different web pages for the result of subject CE

LINK	GRADE
http://results.bput.ac.in/108_RES/100xxxx208.html	F
http://results.bput.ac.in/195_RES/100xxxx208.html	C

Fig-7

With the current system there is no such provision of getting the updated result with the updated SGPA. Hence this represents an unorganized representation of contents related to a particular object spreading over multiple WebPages. With the stated tool (Figure-9) from all the links given in figure-2 to figure-7, the final result and SGPA will be found as shown in figure-8.

BIJU PATNAIK UNIVERSITY OF TECHNOLOGY, ORISSA

Result for B.TECH 1st Semester Examination, 2010-2011

Student Roll No: 100xxxx208
Student Name: xxxxxx xxxx xxxx
College: xxx xxx xxx
Branch: xxx xxx xxx
Published on: 08-Aug-2013

Sl.No.	Subject Code	Subject	Credit	Grade	Date
1	BE2101	BASIC ELECTRONICS	3	C	08-08-13
2	BE2104	MECHANICS	3	D	06-12-13
3	BE2105	PROGRAMMING IN C	3	C	08-08-13
4	BS1101	MATHEMATICS - I	4	F	08-08-13
5	BS1102	PHYSICS - I	3	D	24-06-11
6	HM3101	COMMUNICATIVE ENGLISH	2	C	08-08-12
7	BE7102	WORKSHOP PRACTICE	2	E	24-06-11
8	BE7103	PHYSICS LABORATORY	2	A	24-06-11
9	BE7105	BASIC ELECTRONICS LABORATORY	2	A	24-06-11
10	BE7107	PROGRAMMING IN C LABORATORY	2	A	24-06-11
11	HM7101	COMMUNICATIVE ENGLISH LABORATORY	2	A	24-06-11
		Total Credits:	28	SGPA: 6.00	

Fig-8: [Final Result]

3. BLOCK DIAGRAM FOR THE TOOL

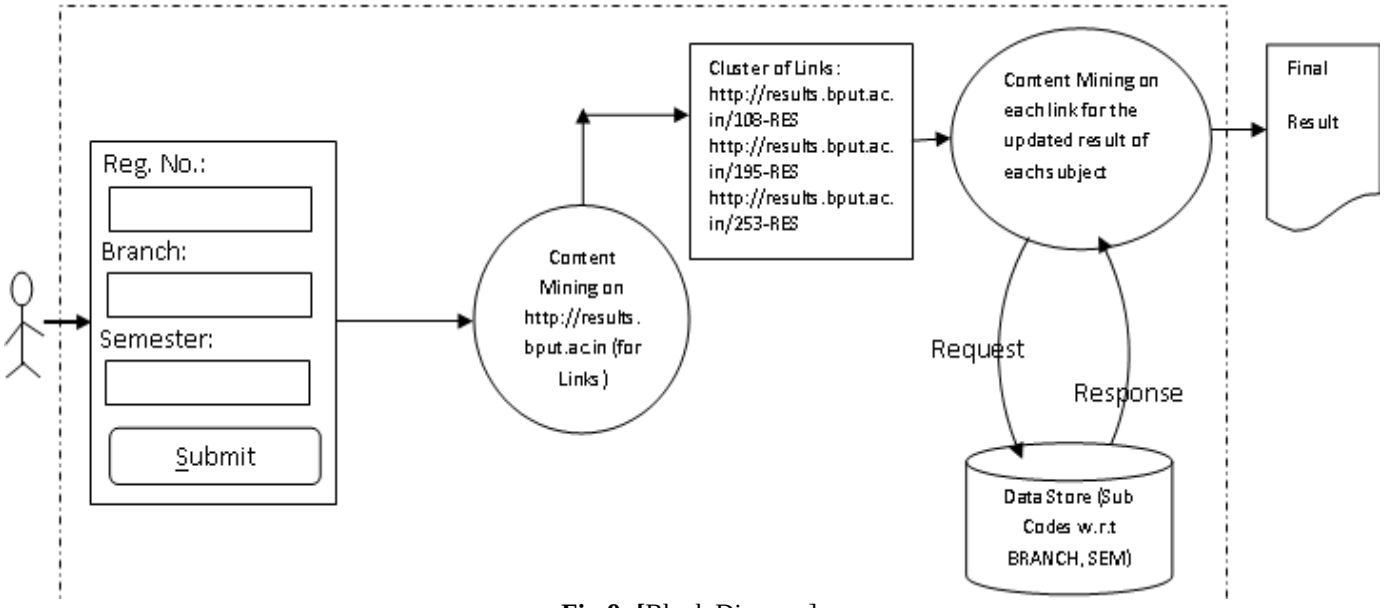


Fig-9: [Block Diagram]

The Working Principle of the stated mining TOOL is as below:

Steps 1 User submits a web form with the input set {Reg. No., Branch, Semester}

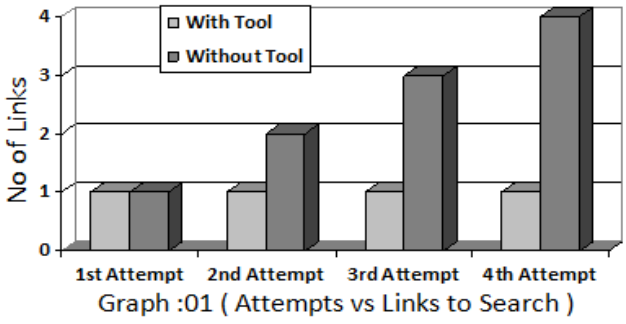
Step 2 The tool applies web content mining on the web page http://results.bput.ac.in/ generating the required cluster of links with respect to input set.

Note: (The cluster of links is dynamic in behavior with respect to input set)

Step 3 The tool applies web content mining on each link from the cluster of links after getting subject codes for the given semester from a data store generating the updated result of each subject.

Step 4 From the result set generated by step 3 the final result gets generated.

4. CONCLUSIONS



Graph :01 (Attempts vs Links to Search)

In this work we had gone through the concepts of web content mining and proposed a tool for generating an organized page from unorganized information spreading over multiple pages. With reference to Graph:01 it can be concluded without the tool, for finding the final result if no of attempts increases, no of links to be searched by user also increases but with the tool, for finding the final result if no of attempts increases, no of links to be searched by user remains one. In future the tool may be modified to find the toughness of a subject with respect to result, grades obtained and no of attempts made by the student in individual subjects.

REFERENCES

[1]. Samia Jones and OmPrakash K. Gupta "WEB DATA MINING : A CASE STUDY" – Communications of IIMA – 2006 Volume 6 Issue 4

[2]. Bing Liu and Kevin Chen-Chuan Chang "EDITORIAL: SPECIAL ISSUE ON WEB CONTENT MINING" - http://www.cs.uic.edu/~liub/publications/editorial.pdf

[3]. Aarti Singh " AGENT BASED FRAMEWORK FOR SEMANTIC WEB CONTENT MINING" – International Journal of Advancements in Technology - Vol. 3 No.2 (April 2012) ISSN 0976-4860 Page 108 - 113

[4]. Federico Michele Facca and Pier Luca Lanzi "RECENT DEVELOPMENTS IN WEB USAGE MINING RESEARCH" - Springer-Verlag Berlin Heidelberg 2003 - DaWaK 2003, LNCS 2737, pp. 140–150, 2003

[5]. Pravin M. Kamde, Dr. Siddu. P. Algur "A SURVEY ON WEB MULTIMEDIA MINING" - The International Journal of Multimedia & Its Applications (IJMA) Vol.3, No.3, August 2011 PP 72 - 84

BIOGRAPHIES

Prof. Binayak Panda has received a bachelor's degree in Computer Science and Engineering from BPUT Odisha in the year 2005. In the year 2010 he has received a master of technology degree in Computer Science and Engineering from BPUT Odisha. Currently he is working as an Asst. Prof. In the Dept. of IT at GIET Gunupur. He has 3 years of industry experiences in the field of Software testing and maintenance. His interested areas of research are Software Engineering and Web Engineering. He is a life time member of ISTE.

Prof. K Murali Gopal has received a bachelor's degree in Computer Science and Engineering from BPUT Odisha in the year 2003. In the year 2007 he has received a master of technology degree in Computer Science and Engineering from BPUT Odisha. Currently he is working as an Assoc. Prof. In the Dept. of CSE at GIET Gunupur. He has 2 years of industry experiences in the field of Software Analysis. His interested areas of research are Software Engineering and Computer Vision. He is a life time member of ISTE.

Prof. Sudhanshu Shekhar Bisoyi has received a bachelor's degree in Information Technology from BPUT Odisha in the year 2005. In the year 2008 he has received a master of technology degree in Computer Science and Engineering from BPUT Odisha. Currently he is working as an Asst. Prof. In the Dept. of CSE at GIET Gunupur. His interested areas of research are Data Mining and Neural Networks. He is a life time member of ISTE.