ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 2 Issue 11, November 2013

A SURVEY ON CRAWLING WEB FORUMS

K.Vidhya
PG Scholar, Department of CSE
SNS College of Engineering

Ms.E.Annal Sheeba Rani M.E.,
Assistant Professor, Department of CSE
SNS College of Engineering

Abstract:

Web forum [1] is an online conversations site where users can hold conversations in the form of posted messages. And then messages are exchanged with others. A discussion is of specific topics and issues. A discussion forum is tree-like in structure: a forum can also contain a number of sub forums. And then a web forum has a different layout and structures are powered by variety of forum software packages. Web forum techniques of iRobot forum crawler [4] it does not maintain a record of previously stored data this is inefficient and time consuming process. Focus (Forum crawler under supervised) technique is supervised one so, this handle only trained data sets and then this is not suitable for dynamic web pages and then the partial tree alignment align the data fields in the pair of data records or contents.

Keywords: iRobot, Focus, Web forum, Partial tree alignment

I.INTRODUCTION

Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. A forum can contain a number of sub forums, each have several topics. Within a forum's topic, each new conversations started is called a thread.

A web crawler is also known as web spider, this is program browses in World Wide Web in a automated manner. Crawlers can also be used for specific type of information and then checking links or validating HTML code.

Web Forum Structure:

A web forum [1] is a tree like or hierarchical structure. A forum can be divided into categories for the relevant conversations and then the posted messages. Under the categories are sub-forums and these sub-forums can be further divided into more sub-forums.

User groups:

A user of the forum can automatically be access to a more privileged user group based on conditions set by the administrator. An anonymous user of the site is commonly known as visitors. Visitors are to granted access to all functions that do not require breach privacy. A guest can usually view the contents and then the posted messages of the forum.

Moderators

The moderators are called visitors or users of the forum who are granted access to the posted messages and threads of all members for the purpose of moderating conversations and also keeping the forum clean Moderators also answer users' concerns about the forum, general questions, as well as take action to specific complaints in the conversations.

Posts:

A post is a user will hold conversations to submit a message enclosed into a block containing the user's details and the date and time it was submitted. Posts have a limit usually measured in the characters. To have a message of minimum length of 10 characters. There is always an upper limit most boards have it at either 10,000, or 20,000 characters.

Thread

A thread is a collection of posts, conversations of the messages usually displayed from oldest to latest, A thread can contain any number of posts, including multiple posts from the same members, even if they are one after the other.

II.LITERATURE SURVEY

1) Web data extraction based on partial tree alignment:

DEPTA (Data extraction based on partial tree alignment) [2]
This method consists of two steps:

1) Identifying the individual records in a page of a web forum of the posted messages or the conversations.

2) Aligning and extracting the data items from the Identified records of the web forum.

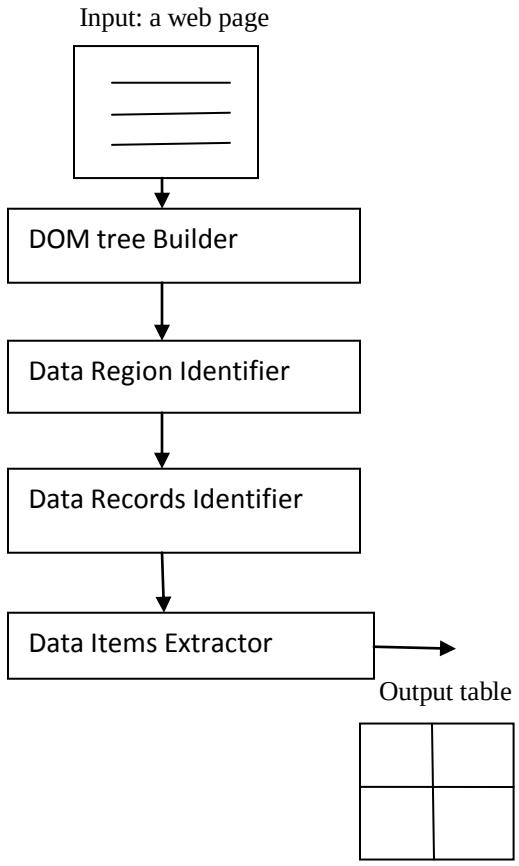

Fig:1 Architecture diagram for DEPTA

The Partial tree alignment approach:

- Choose a seed tree: A seed tree , denoted by Ts, is picked with the maximum number of data items.
- Tree matching:
- For each unmatched tree Ti (i≠s),
  - Match Ts and Tr

-Each pair of matched nodes are linked (aligned)

-For each unmatched node nj in Ti do

Expand Ts by inserting n into Ts if a position for insertion can be uniquely identified and then determined in Ts.

The expanded seed tree Ts is used to the subsequent matching.

Disadvantages:

- Cost overhead
- Tree alignment techniques have been used. So traversal time was high.

2) To Extract Structured Data from Web Forums

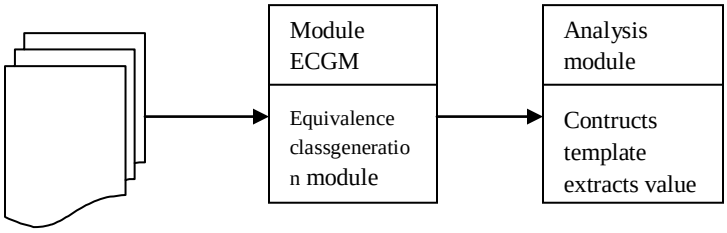

Fig 2: General description EXALG

EXALG

- Full automatic extraction [2] of structured data from web forums of conversations.
- In order to post the complex queries over data in the web forum.
- Avoid human input

Challenges of EXALG:

- Distinguish template from data and then deduce the scheme for encoded information.
- Complex schemas and arbitrary levels are nested.

Disadvantages:

- Excess time is needed
- Performance is limited.

3) An Intelligent crawler for web forums:

iRobot forum crawler[4] is used, which will crawl the forum content, it does not deal with the frequent thread updation in the web forum.

iRobot forum crawler is does not maintain a record of previously stored data. It is a tree like traversal.

Disadvantages:

- More time consuming process.
- No clear understanding of page identification is carried out.

4) Incorporating site level knowledge a list wise strategy:

- Distinguish index and post pages
- Concatenate pages to list by following paginations

Page Layout Clustering

- Forum pages are based on a database and then the template.

- Layout is robust to describe template

Layout can be characterized by the HTML elements in different DOM paths (e.g. repetitive patterns)

Indentify Index & Post Nodes:

A SVM based classifier is used.

- It is a site independent

- Node classification is robust that page

  - Robust to noise on individual pages

5) Exploring traversal strategy for web forum crawling:

Traversal strategy:

- Skeleton link identification
- Page-flipping link detection

Skeleton link identification:

The most important [5] links supporting the structure of a forum site.

Two characteristics:

- Skeleton links should point to those vertices and then containing the valuable and informational pages.
- A link is not a skeleton link if it points to many duplicate pages or valueless pages.

The content-based near-duplicated detection algorithm is employed.

Two Criteria:

- Coverage
- Informativeness

Page-Flipping Link Detection:

- Page-flipping link is a one kind of loop-back links.
- Not all the loop-back links are page-flipping links.

- Connectivity is higher than the average score are selected as the page-flipping links.

$$\text{Connectivity}=\frac{\sum_{\{A,B\}} \text{Path}(A,B).\text{Path}(B,A)}{\sum_{\{A,B\}} \text{Path}(A,B)}$$

III.CONCLUSION

In this paper, the five crucial techniques for mitigating the web forums were discussed: (1) DEPTA technique (2) EXALG algorithm (3) iRobot forum crawler technique (4) Template-independent approach (5) Traversal strategy technique Under this survey, it is concluded that all the methods uses several terminologies for web forum in the presence of an adversary and the successful conveyance of any data. In this paper, the technique used is crucially adapted for time consuming, performance and then accuracy.

IV.REFERENCES

1) Internet Forum, http://en.wikipedia.org/wiki/Internet_forum, 2012.

2) Y. Zhai and B. Liu, “Structured Data Extraction from the Web based on Partial Tree Alignment,”

IEEE Trans. Knowledge Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

3) J.-M. Yang, R. Cai, Y. Wang, J. Zhu, L. Zhang, and W.-Y. Ma, "Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums," Proc. 18th Int'l Conf. World Wide Web, pp. 181-190, 2009.
4) R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang, "iRobot: An Intelligent Crawler for Web Forums," Proc. 17th Int'l Conf. World Wide Web, pp. 447-456, 2008.
5) Y. Wang, J.-M. Yang, W. Lai, R. Cai, L. Zhang, and W.-Y. Ma,"Exploring Traversal Strategy for Web Forum Crawling," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development inInformation Retrieval, pp. 459-466, 2008.
1) Internet Forum, http://en.wikipedia.org/wiki/Web mining, 2012.

AUTHORS:

Vidhya K received BE degree in computer science and Engineering from Kalaignar Karunanidhi Institute of Technology, Coimbatore . She is currently pursuing her ME degree in SNS College of Engineering, Coimbatore. Her Research interest includes data mining and network security.

Annal Sheeba Rani E received BE degree in computer science and Engineering from Arulmigu Kalasaligam College of Engineering and ME degree in computer science and Engineering from Anna University Chennai and Presently she is working as Assistant Professor in Department of Computer science and Engineering, SNS College of Engineering, Coimbatore Her Research interests includes image processing and data mining.