ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

A Survey on Web Forum Crawling Techniques

T. Mahara Jothi, K.Thirumoorthy

PG Scholar, Dept of computer science and Engineering, Mepco Schlenk Engineering College,Sivakasi, India.

Assistant Professor, Dept of computer science and Engineering, Mepco Schlenk Engineering College,Sivakasi, India.

Abstract This paper focuses on the study of web forum crawling problem which is an important task in web applications such as web mining and search engines. Due to the richness of the information contributed by millions of internet users every day, web forum sites have become precious deposits of information on the web. As a result, mining knowledge from forum sites has become more important and more significant. However, forum sites exist in different layouts or styles and they are powered by different software packages which makes forum crawling, a tedious task. In addition, large amount of duplicate pages and uninformative pages on forum sites also makes forum crawling task inefficient. In this paper, various forum crawling techniques and their comparisons has been discussed.

Index Terms – Forum crawling, page-flipping links, sitemap, thread page.

I.INTRODUCTION

Internet forum (also known as web forum) is an online discussion place on a website. Forum sites allow it's user to request and exchange information among them. In addition, the forum sites allow users to view forum postings and to post messages in it. When posting in a forum, the users can create new topics (or "threads") or post replies within existing threads. Web forums are almost available for all kinds of topics. Examples include software support, help for webmasters, and programming discussions, sports, entertainment, games, technical discussions etc.

The Internet forums are comprised of user-generated content (UGC), so they continue to grow as long as users visit post messages in the forum sites. Due to this constant growth, they have become precious deposits of information on the Web. Thus the researchers are increasingly interested in mining knowledge from the forum sites. To extract knowledge from the forum sites, their content should be downloaded first. However, crawling data from the forum sites is not a trivial task. Generic crawlers [10] that adopt breadth-first traversal strategy are not suitable for Internet forum crawling problem mainly due to the following reasons: 1) Duplicate links, 2) Uninformative pages, 3) Page-flipping links and 4) Entry URL discovery.

A forum site usually has many duplicate links that point to a common page but with different URLs such as "view by date" and it also has many uninformative pages such as login control, advertisements etc. Following these links, a crawler will crawl many uninformative pages thus making it inconsistent for forum crawling.

Usually a long forum boards or threads are divided into multiple pages that are linked by page-flipping links. Those links should be preserved to facilitate tasks such as page wrapping and content indexing etc. And also, for efficient forum crawling process, the crawling should start with the entry URL of the forum. In addition, for a forum crawling technique to yield high performance, forum entry URL discovery is required.

This paper gives the overview of various forum crawling techniques like iRobot, FOCUS, ILSK, BFC, WeRe, ETS. The rest of the paper is organized as follows: The structure of forums is explained in section II. In section III, each technique is discussed in detail. Section IV provides the comparison between each technique. Section V concludes the paper.

II. WEB FORUM STRUCTURE

An internet forum [9] is hierarchical or tree-like in structure: each forum contains a number of sub-forums (also known as boards), each of which may have several topics. Within a forum's topic, each new discussion started is called a thread, and can be replied to by as many people. Each page in forum site may have its own layouts. Based on their layout structure, the pages in forum sites are classified into four categories:

- Entry page: The home page of the forum site which contains a list of boards.
- Index page: An index page contains table-like structure, where each row in the table contains information of a board or thread.
- Thread page: A thread page contains a list of users' posts.
- Other pages like login control, about us, user profile pages, etc.

Every forum site has similar navigation paths though they differs in layout and styles structure. The users' of forum site usually follows the navigation path as given below:

Entry page → Index page → Thread page

The hierarchical page and link structure of forum site is illustrated in figure 1.

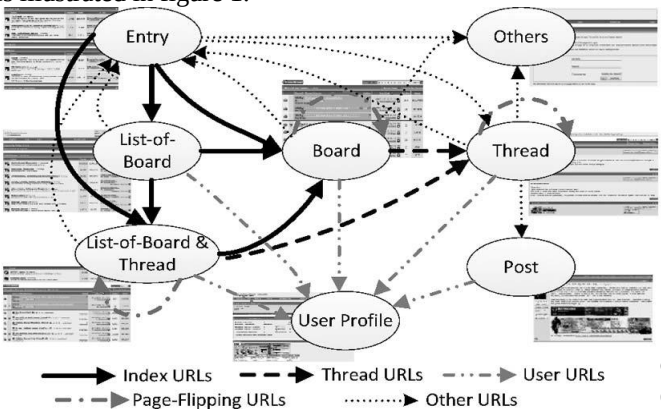


Fig. 1 Structure of forum sites

categories: index URLs, thread URLs, page-flipping URLs, user URLs, and other URLs. As shown in figure 1, the links that connects entry page and index page or links that connects two index pages are known as index URLs. Thread URLs are used to connect index pages and thread pages. Page-flipping links are used to connect multiple pages belonging to the same board or thread. The board or thread pages are connected to user profile pages by user URLs. The links that are used to connect board or entry or thread pages to pages like about us, FAQs are referred here as other URLs.

To make forum crawling an efficient process, there is a need to extract the above type of URLs from the forum site. Among the extracted URLs, the index URLs, thread URLs, page-flipping URLs are then used for extracting user posts' from forum sites. The user URLs and other URLs are discarded since those URLs point to uninformative pages. In addition, the entry URL that points to home page of forum needs to be identified for efficient forum crawling process.

III. FORUM CRAWLING TECHNIQUES

The forum crawling techniques like iRobot, FOCUS, Board Forum Crawling, Incorporation of site level knowledge to extract structured data from forums, WeRe, ETS are discussed in this section.

A. Board Forum Crawling (BFC)

Usually to visit a post in a forum site, human starts from the home page, then enters into the board, and then to find the post. This process hints that the forums are organized structurally. Board forum crawling [6] method exploits such organized characteristics of forum sites and it simulates the human behavior of visiting web pages for extracting data form forum sites.

This method starts its crawling process from home page, then it extracts the links of board pages from home page (also known as board page seeds), then the subsequent board page links or page-flipping links (also known as subsequent board page seeds) are extracted from each board, then user posts are extracted from each subsequent board page seed.

Given the homepage of the forum site, it performs crawling process as follows:

1. *First it e*xtracts board page seeds from homepage
2. Then, for each board page seed, link queue of all subsequent board pages in the same board is created.
3. For each queue, it downloads each page in the queue, identify whether it is exactly a board page and extract links of post pages from the board page.
4. Then it creates the whole link index of all post pages in all board pages.
5. Finally, post pages linked by the whole index obtained in step 3 are downloaded.

Board forum crawling technique is simple forum crawling technique, but it doesn't deal with entry URL discovery problem and it provides little support for duplicate link detection while comparing with other techniques.

B. iRobot: An Intelligent Web Forum Crawler

This approach [4] automatically understands the content and structure of each forum sites and then decides how to traverse to different pages in forum sites. To find out such traversal paths, it first automatically re-build the sitemap [5] of the target web forum and then it selects the optimal traversal paths which only traverses informative pages and skip invalid and duplicate pages. Figure 2 shows the architecture of iRobot system.

iRobot system consist of two major parts: 1) offline sitemap reconstructing and traversal path selection 2) online crawling. The offline sitemap reconstructing and traversal path selection part involves four major steps:

1) Repetitive region-based clustering
2) URL-based sub-clustering
3) Informativeness estimation
4) Traversal path selection

The offline part starts by randomly sampling some few pages form the target forum site. The sampled pages are then given as input to repetitive region-based clustering step. Then the following steps are carried out:

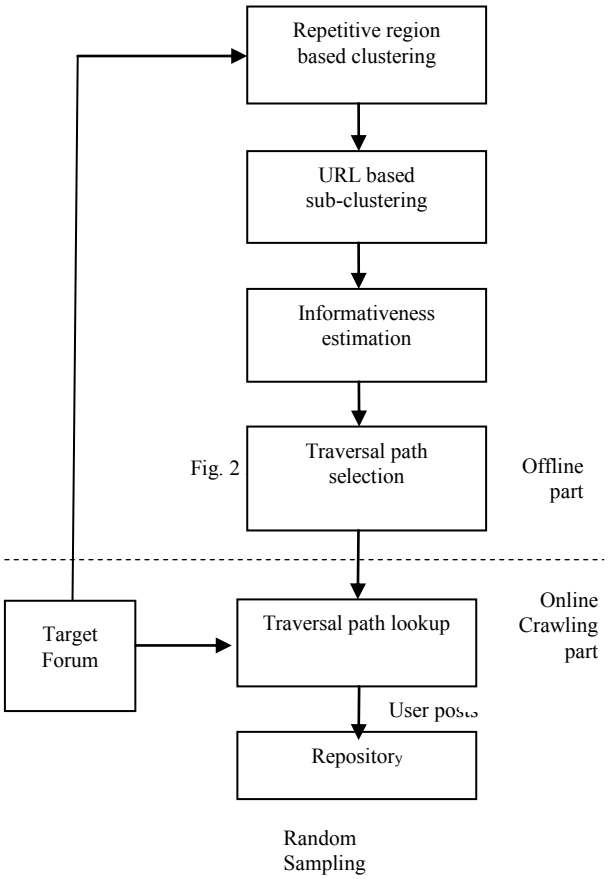


Fig. 2

A repetitive pattern on a web page is a block area containing multiple data records in a uniform formation. In this step, first the pages having similar layout (repetitive regions) are clustered from the sampled pages. Once the repetitive regions are identified, then following steps are carried out:

- First, the repetitive pattern (an abstract representation of all records in repetitive region) is generated for each repetitive region in every page.
- Feature description is then created for every page by recording the number of times the repetitive pattern occurs in that page.

ii) URL-Based Sub-Clustering

In this step, each layout clusters that were found in previous step are further split into subsets by grouping those pages with similar URL formats. The similarity between any two URL addresses are found out based on the following two assumptions:

1) URLs having the same number and the same order of the paths are said to be similar.
2) URLs are said to be similar, if both have same parameters of keys.

After the subset clustering, each cluster is finally represented with a URL pattern (a sequence of regular expressions generated for every segment of paths and parameters in URL). Each obtained subset cluster is then taken as a vertex of the sitemap. Finally this step connects various vertices of the sitemap with each other.

iii) Informativeness Estimation

The pages that are having more valuable information are found out in this step. The pages of forum sites are said to be informative if it should satisfy following assumptions:

1) Pages in the large cluster with high probability are valuable.
2) A valuable page usually has relatively large file size.
3) The semantic diversity of each page in cluster is also used to find informativeness of that page.

Based on the above three assumptions, the infomativeness measure is calculated for each page. The pages having high informativeness value are selected and the other remaining pages are discarded in this step.

iv) Traversal Path Selection

The traversal path selection consists of two major parts: 1) cleaning the sitemap 2) optimal traversal path selection. The sitemap is automatically cleaned by removing most useless vertices and arcs in it, by following the heuristics below:

1) Vertices with low quantity of information are dropped.
2) For a layout cluster, containing several vertices, reserving one representative is enough, as the others are prone to be duplicates.
3) For each vertex, arcs of self-linking are removed except those whose anchor text is a digital string or some particular strings such as "next".

Finally, the optimum traversal path that traverses all the survived vertices with minimum cost is found out. At last, in online crawling step, a downloaded page is first classified into one of the vertex on the sitemap. Then for an out link from that page, the traversal path lookup step further find out its URL pattern and location on that page and finally determine how to follow it by looking up in the lookup table. And for each link in the link table, it decides whether that link should be added to the crawling queue based on the list of traversal paths.

The main advantage of iRobot system is that it requires less human intervention and it provides support for duplicate links and uninformative pages removal. But its sampling strategy and informativeness estimation is not robust.

C. Incorporation of Site-Level Knowledge to Extract Structured Data from Forum Sites (ISLK)

This approach [3] focuses on extraction of structured data from forum sites such as users' post title, post content, post time, and post author. It incorporates both page-level knowledge and site-level knowledge and employs markov logic networks to draw the joint inference from both the knowledge's to extract the structured data. For this, the page-level knowledge such as link to user profile, timestamp existence, order of timestamp (ascending or descending) etc are learnt as features from individual pages of forum sites. The site-level knowledge representing linkages among different pages in forum sites and the interrelationships of pages belonging to same page type are obtained by re-constructing the sitemap of forum sites. The flowchart of the method is illustrated in figure 3. It consists of three main parts: 1) offline sitemap recovering, 2) feature extraction 3) joint inference of pages having same template.

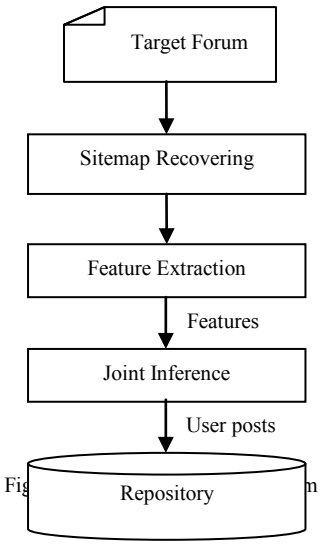


Fig m

This step is similar to the offline sitemap recovering step of iRobot [4] technique.

ii) Features Extraction

In this part, the DOM tree is first constructed from the HTML content of the forum page. Once the DOM tree is constructed, the following three kinds of features are extracted: 1) *Inner-page features* includes the features that leverage the relations among the elements inside each page such as inclusion relation among elements, timestamp existence, order of timestamp, size and location of each elements in page. 2) *Inter-template features* that are generated based on the site-level knowledge includes existence of link to user profile, existence of link to users' posts etc. 3) *Inter-page features* includes existence of text elements having similar DOM path and tag attributes, existence of hyperlink elements with similar DOM path and tag attributes, and existence of inner elements with similar DOM path and tag attributes are extracted.

iii) Joint Inference of Pages having Same Template

In this step, the Markov Logic Networks are used to combine the features extracted in previous step efficiently to draw the joint inference to extract data such as post title, post author, etc. The rules / formulas are defined for extracting data from list pages and post pages separately. Based on those rules the data like list records, list title are extracted from list pages and data like post records, post author, post time, post content are extracted from post pages respectively.

D. FOCUS (Forum Crawler Under Supervision)

This is a supervised web forum crawler. This approach automatically learns the regular expression patterns of navigation paths in order to crawl the relevant forum contents (users' posts) with minimal overhead. The architecture of FOCUS [1] system is illustrated in figure 4. It consists of two major parts: learning part and the online crawling part. The learning part first learns ITF (Index-Thread-Page-Flipping) regexes from extracted index, thread, page-flipping URLs. The online crawling part then tries to crawl all thread pages that matches the learned ITF regexes. The learning part is composed of four modules: 1) Entry URL discovery, 2) Index/Thread URL detection, 3) Page-Flipping URL detection, and 4) ITF regexes learning. *i)Entry URL Discovery :*To start crawling process, entry URL of forum is required. Works like BFC assumed that forum entry URL is given. But for large-scale crawling, manual entry of forum entry URL is not practical, so it is necessary to find out the entry URL of the forum automatically. The Entry URL of forum site is extracted based on the following observations:

1) Almost every page in forum site contains a link back to its entry page.
2) The home page of the site hosting a forum contains an entry URL of forum.
3) Index URL should not be entry URL.
4) An entry page has more index URLs.

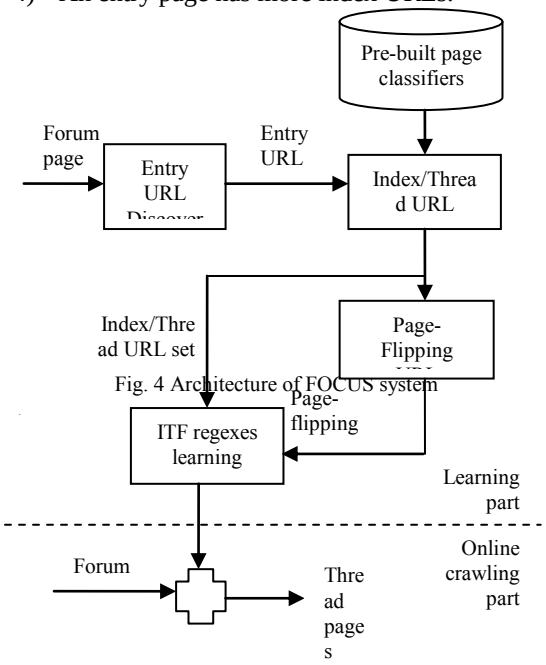


Fig. 4 Architecture of FOCUS system

This module identifies the index/thread URLs present in the forum sites. Both index and thread pages have their own layouts. For example, index/thread URLs are grouped and placed in the same table column position. Such URLs are grouped based on their locations using partial DOM tree alignment method. Once such groups are constructed, URL group with longest anchor text is extracted. Then for each URL in that URL group, it's page type need to be identified. For that, the system built index and thread page classifiers by extracting features based on the layout characteristics of index and thread pages. Once the page type is identified, a majority voting method is used to determine the URL type.

ii) Page - flipping URL detection

The system then identifies the page-flipping URL that connects multiple pages of a board/thread. The URLs satisfying the following properties are extracted as page-flipping URLs

1) The anchor text of such URLs contains either digit sequence such as 1, 2, 3 or text such as "next", "last".
2) Those URLs appear at the same location on DOM trees.
3) The destination pages of those URLs have similar layout.

4) Some page-flipping URLs appearing in the source and destination pages have the same anchor text but different URL strings.

iii) ITF Regexes Learning

In this module, URL regular expression patterns are learned from index, thread and page-flipping URLs. It starts with finding general URL patterns, it finds out more specific URL patterns matching a set of URLs. Then each specific pattern is further refined into more specific patterns and the patterns are refined recursively until no more patterns can be refined. Each such learned ITF regexes contains three elements namely: page type, URL type and URL pattern.

Once the ITF regexes are learned, then online crawling is performed as follows: starting from the entry URL, it follows all URLs that are matched with any learned ITF regex.

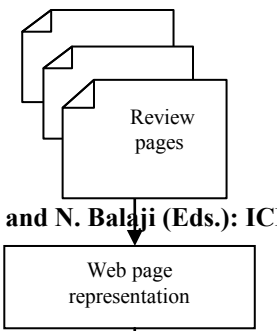
Entry URL discovery, elimination of duplicate links and uninformative pages by learning ITF regexes and providing support for page-flipping links detection are some advantages of this system.

E. *Automatically Extracting User Reviews from Forum Sites (WeRe)*

The WeRe (Web Review) extraction system [7] automatically extracts the review records (users' posts) from forum pages using level-weighted tree similarity algorithm. Then the review contents are extracted from the review records by measuring the consistency of each review record nodes in DOM tree and by extracting the minimum subtree that contains the pure review content based on the measured node consistency values. In addition, this approach provides two extraction choices: direct extraction and wrapper-based extraction. The architecture of WeRe system is illustrated in figure 5. The WeRe System consists of four main modules: 1) Web page representation, 2) Review record extraction, 3) Review content extraction, 4) Wrapper-based extraction.

i) Web Page Representation

Fig. 5 Architecture of WeRe system



Once a review page is obtained as input, this module parses a review page into a DOM tree. Then the Visual information is extracted from review pages and is attached to the nodes of the DOM tree. Visual information include position of a node in DOM tree, width and height of the rectangle that node occupies in web page, the fonts of texts of the nodes including font size, color and style.

ii) Review Record Extraction

This process involves three steps: 1) first, the review region is detected in the DOM tree, 2) then, the noise is removed from review region, 3) finally, the boundaries of review records is extracted. The review region is detected from DOM tree by extracting visual features from review pages. Usually the review records present in one review page have same template, same semantics, and are more similar. Such review records are obtained from DOM tree, by extracting subtrees of review region having the same semantics using level-weighted tree similarity algorithm. Then the unwanted noise information present in the review region is removed. At last, the boundaries of the review records are detected by constructing global similarity relationship graph.

iii) Review Content Extraction

The review records may contain information like author details, post time and actual post, etc. This process first extracts the review region containing review contents. Then it extracts review contents of different review records and then aligning it into a supertree. Then the consistency of each node and consistency of subtrees in constructed supertree are computed. Finally, the review contents are extracted based on computed consistency values.

iv) Wrapper-Based Extraction

Given a review page, the review records and review contents are extracted with the help of above three modules. In addition, this approach also constructs the wrapper for the extraction of review contents. It generates

two wrappers one for review record extraction and other for review content extraction are constructed by extracting visual features from sample pages as discussed in web page representation module. Finally, the review records and review contents are extracted with the help of generated wrappers.

The WeRe system extracts the review content information efficiently. This technique directly starts with review page, thus it doesn't provide support for entry URL detection, page-flipping links detection which makes it little inconsistent for forum crawling process.

F. Exploring Traversal Strategy for Web Forum Crawling (ETS)

This approach [8] tries to explore the traversal strategy to direct the forum crawling process. This approach identifies the traversal strategy by re-constructing the sitemap of forum site as in iRobot [4] and then by identifying the skeleton links and detecting the page-flipping links present in the forum site. This system consists of two major steps: 1) sitemap recovering, 2) exploring traversal strategy. The architecture of this system is shown in figure 6.

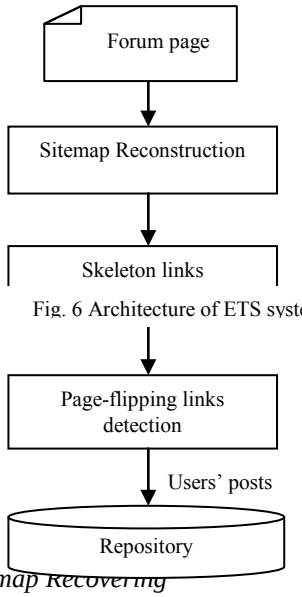


Fig. 6 Architecture of ETS system

i) Sitemap Recovering

This step is similar to the offline sitemap recovering step of iRobot [4] technique.

ii) Exploring Traversal Strategy

This step involves the detection of skeleton links and page-flipping links. The skeleton links are identified form the reconstructed sitemap by traversing from vertices of sitemap in top levels to vertices in deeper levels. During this traversal process, the processed vertices are put into one set and unprocessed vertices are put into another set. Then for each vertex in unprocessed vertex set, all the links from the vertices in processed vertex set pointing to that vertex in unprocessed vertex set are extracted as skeleton links.

Usually the page-flipping link have the following characteristic: if there is a path from a forum page A to another page B, then there must be a path from B to A. After skeleton link detection, this step extracts the link satisfying the above characteristic as page-flipping link. For that, it first extracts the outgoing links present in that page. Next, the connectivity measure is calculated for each extracted links. Finally, the links having highest connectivity measure values are extracted as page-flipping links.

The ETS system efficiently extracts the users' posts by detecting the page-flipping links and by removing the duplicate pages present in forum sites. But this system doesn't provide support for extracting the entry URL of the forum site.

IV. A Comparative Analysis of Forum Crawling Techniques

In this section, the comparison of forum crawling techniques is discussed. Some criteria that were discussed in previous work [1] like extraction rule type, features used, learning algorithm etc. are used to compare the above discussed forum crawling techniques. The comparison between the forum crawling techniques is shown in table 1.

The forum crawling technique like FOCUS uses extraction rules that are represented as regular expressions for eliminating duplicate links. Some techniques like ISLK uses extraction rules expressed using first order logic for extracting data from forum sites efficiently.

Techniques like FOCUS, ISLK, iRobot uses DOM tree to extract features since the data to be extracted are often located in the same path of DOM tree, thus using DOM tree makes data extraction process much easier. And the technique like BFC uses HTML tags to extract data from forum sites.

Learning algorithm plays an important role in extracting data from web pages. For example, techniques like FOCUS, ISLK, iRobot, WeRe, ETS uses partial DOM tree alignment technique to mine data records from forum sites. BFC uses clustering technique to mine board pages and subsequent board pages from forum sites.

In addition, providing support for entry URL and duplicate links detection are also required for efficient forum crawling process. Techniques such as iRobot, ILSK, WeRe, ETS provides support for duplicate links detection but they doesn't provide support for entry URL discovery. But FOCUS provides support for both entry URL discovery and duplicate links detection.

Each forum sites may have its own layout or template structure. The variation in layout structure may cause difficulties in extracting data from forum sites. So, there is a need to deal with such template variations. ILSK technique uses disjunctive rule support and the BFC, ETS techniques use clustering technique to deal with such template variations.

Table 1: Comparison of forum crawling techniques

S.no.	Criterion	Forum Crawling Techniques					
		BFC	**iRobot**	**ILSK**	**FOCUS**	**WeRe**	**ETS**
1.	Extraction rule type	None	URL patterns	First order logic	Regular expression	Similarity Measurement	URL patterns
2.	Features used	HTML tags	DOM tree	DOM tree	DOM tree	DOM tree	DOM tree
3.	Learning algorithm	Clustering	Partial tree alignment	Partial tree alignment	Partial tree alignment	Partial tree alignment	Partial tree alignment
4.	Entry URL discovery	Manual entry	Manual entry	Manual entry	Automatically identified	Manual entry	Manual entry
5.	Support for duplicate link detection	Medium	High	High	High	Medium	High
6.	Template variation	Clustering	Clustering	Disjunctive rules	Regular expressions	-	Clustering

From the above comparison it is clear that FOCUS outperforms other techniques in terms of effectiveness, since FOCUS provides support for entry URL discovery, and also deals with duplicate links detection and uninformative pages detection with the help of ITF regexes and it also provides support for page-flipping links detection.

V. CONCLUSION

In this paper, we presented a short survey of forum crawling techniques. The techniques like iRobot, Board Forum Crawling, Incorporation of site-level knowledge, and FOCUS that is used to extract structured data from forum sites are discussed in brief. We compare those techniques based on some set of criteria like extraction type, extraction rule type, learning algorithm, template variation, support for entry URL discovery, and support for duplicate links detection. From the survey and comparison, it is clear that FOCUS outperforms those techniques in terms of effectiveness and coverage by providing good support for entry URL discovery, duplicate links detection, support to deal with template variation.

REFERENCES

[1] J. Jiang, X. Song, and N. Yu, "FoCUS: Learning to Crawl Web Forums," *IEEE Trans. Knowledge and Data Engg*, pp. 1293-1306, 2013.

[2] M. Kayed, C.-H. Chang, M.R. Girgis, and K. Shaalan, "A Survey of Web Information Extraction Systems," *IEEE Trans. Knowledge and Data Engg*, vol.18, pp. 1411-1428, 2006.

[3] 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 Intellignet Crawler for Web Forums," *Proc. 17th Int'l Conf. World Wide Web*, pp.447-456, 2008.

[5] U.Schonfeld, and N.Shivakumar, "Sitemaps: Above and Beyond the Crawl of Duty," *Proc. 18th Int'l Conf. World Wide Web*, pp.991-1000, 2009.

[6] Y. Guo, K. Li, K. Zhang, and G. Zhang, "Board Forum Crawling: A Web Crawling Method for Web Forum," *Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence*, pp. 475-478, 2006.

[7] W.Liu, H.Yan, and J.Xiao, "Automatically extracting user reviews from forum sites," *Elsevier: Computers and Mathematics with Applications*, pp. 2779-2792, 2011.

[8]

[9] Y. Wang, J.-M Yang, W.Lai, R.Cai, W.-Y. Ma, and L. Zhang, "Exploring Traversal Strategy for Web Forum Crawling," *Proc. 31 st Ann. Int'l ACM SIGIR Conf. Research and Development in Information*, pp. 459-466, 2008.

[10] Internet Forum, http://en.wikipedia.org/wiki/Internet_forum, 2014.

[11] Web Crawler, http://en.wikipedia.org/wiki/Web_Crawler, 2014.