Article can be accessed online at http://www.publishingindia.com

Web Forum Crawling using Index Thread Page Flipping Algorithm

A. Anny Leema*, P. Iswarya**

Abstract

Internet forums are important platforms where users can send request and exchange information from different sources. The issue in existing system is the URL type recognition problem which consists of duplicate links and uninformative pages. Index Thread Page Flipping Algorithm (ITF) is used to overcome this issue. URL layout and page layout are used to recognise whether the URL link is valid or invalid.

In this project (Phase-I), "Web Forum Crawling using Index Thread Page Flipping Algorithm" is provided that finds whether the links are valid or invalid. The goal is to crawl relevant content. The Internet forums will have the URL type recognition problem. It learns to get the correct path or URL by using regular expression patterns and with created training sets from page type classifiers.

The modules implemented are user interface design module, page flipping module, entry URL discovery module, index/thread URL detection module, generic crawler module. In the user interface design module to connect with a server, user must give their user name and password. In the page flipping module, a long forum is divided into more pages which are linked by page-flipping links. Generic crawlers process each page individually and ignore the relationships between such pages. In the entry URL discovery module entry URL should be specified to perform the process. Some rules are defined to find the entry URL. In the index and thread URL detection module, index URL and thread URL are identified by their URL pattern. In the generic crawler module, given a forum it enters into the thread page and it performs crawling where it avoids the duplicate links and page flipping links.

The front end for all the modules in the project (Phase-I) is designed using eclipse and the backend is designed using SQL server 2005. The two modules in the project (Phase-I) are implemented using Java Servlet, JSP and the code behind is written using Java. The main feature of this project (Phase-I) is to save the bandwidth and time.

Keywords: Forum Crawling, Index Url, Thread Url, Page Flipping Url

Introduction

This chapter deals with the general introduction of the project (Phase- I), existing system and the proposed system of this project (Phase- I). Web forum has now become an important data source of many web applications; while forum crawling is still a challenging task due to complex in-site link structures and login controls of most forum sites.Without carefully selecting the traversal path, a generic crawler usually downloads many duplicate and invalid pages from forums, and thus wastes both the precious bandwidth and the limited storage space.

The skeleton links instruct the crawler to crawl only valuable pages and meanwhile avoid duplicate and uninformative ones; and the page-flipping links tell the crawler how to completely download a long discussion thread which is usually shown in multiple pages in Web forums. The extensive experimental results on several forums show encouraging performance of the approach. Following the discovered traversal strategy, forum crawler can archive more informative pages in comparison with previous related work and a commercial generic crawler.

* Assistant Professor, Department of Computer Applications, B.S. Abdur Rahman University, Chennai, Tamil Nadu, India.
** Department of Computer Applications, B.S. Abdur Rahman University, Chennai, Tamil Nadu, India.
Email: Iswaryaponnuswamy007@gmail.com

Related Work

(i) *iRobot: An Intellegent Crawler for Web Forums*, R. Cai, J.-M. Yang, W. Lai, Y. Wang, and L. Zhang

In this paper iRobot has intelligence to understand the content and the structure of a forum site. It decides how to choose traversal paths among different kinds of pages.

(ii) *Detecting Near Duplicates for Web Crawling*, G. S. Manku, A. Jain, and A. D. Sarma

This paper proposes the two documents which are identical in content but differ in a small portion of the document such as advertisement etc.

Suppose newly crawled page is near duplicate of already crawled page then it ignores the link.

(iii) *Extracting and Ranking Product Features in Opinion Documents*, L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain

Opinion mining is to extract people's opinions on features of an entity. For example, "I love the GPS function of Motorola Droid" expresses a positive opinion on the "GPS function" of the Motorola phone.

(iv) *Incorporating site-level Knowledge to Extract Structured data from web forums*. J.M. Yang *et al.*

The target is to find a solution as general as possible to extract structured data like post-title, post-author, post-time, post-content from the site.

Disadvantages of the Existing System

The major disadvantages of existing System are as follows.

- It does not find duplicate links and uninformative pages.
- Page flipping links will have multiple pages. It will process each page individually and ignore the relationship.
- URL type Recognition Problem.

Proposed System

In proposed system the problem is overcome by using ITF Algorithm. It recognizes the Index URL, Thread URL and page flipping URL using page type classifiers. It eliminates the duplicate links. Forums exist in many different layouts (or) styles and are powered by variety of forum software packages, but they always have implicit navigation path to lead users from entry pages to thread pages.

Advantages of the Proposed System

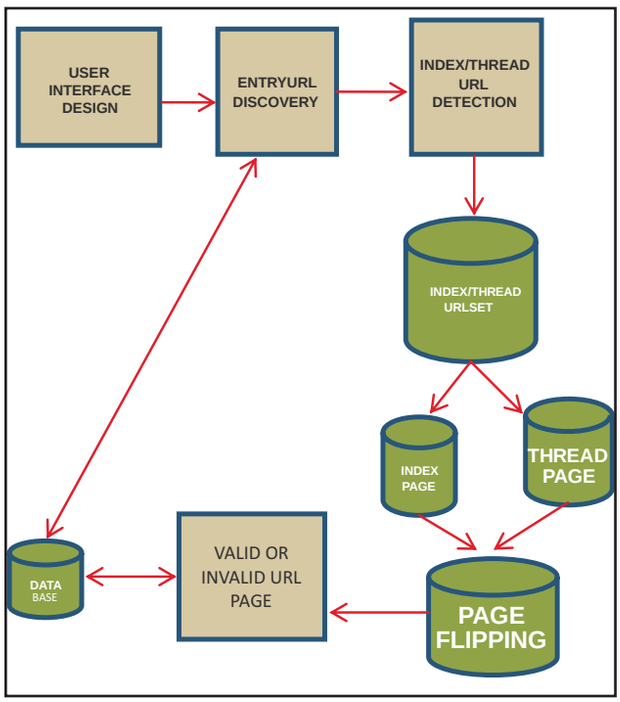
The advantages of the proposed system are

- Eliminates duplicate links.
- User friendly
- Design an effective forum entry URL discovery method.
- The learned patterns are effective and the resulting crawler is efficient.

Architectural Design

The system architecture defines the structure of the developed system, which comprises different components or modules, their externally visible properties and the relationships among them. Figure 1 explains the overall architecture diagram for this project (Phase- I).

Figure 1: Architectural Diagram



Detailed Design

Detailed design will explain the software components in detail. This will help in the implementation of the system. Moreover, this will guide the further chances in the system to satisfy the future requirements.

The detailed design explains every module in detail.

- User Interface Design
- Page Flipping
- Entry URL Discovery
- Index and Thread URL Detection
- Generic Crawler

Module Descriptions

User Interface Design

In the user interface design module, to connect with a server user must give their user name and password. User can login directly if they have already registered otherwise they have to register their details username, password, email and country into the server. Once the user is logged in, the login token may be used to track what action the user has taken when connected to site.

Page Flipping

In the page flipping module, a long forum is divided in more pages which are linked by page flipping links. This module helps to avoid the number of pages. Once a user enters a forum page there will not be any page flipping links. All the pages, that is, the number of user posts, are on a single long forum thread.

Entry URL Discovery Module

In the entry URL discovery module, entry URL should be specified to perform the process. Some rules are defined to find the entry URL. It tries to find the following keywords ending with "/" in a URL: forum, board, community, and discuss. If a keyword is found, the path from the URL host to this keyword is extracted as its entry URL.

Index and Thread URL Detection

In the **index and thread URL detection** module, index URL and thread URL are identified by their URL pattern. Index URL and thread URL will contain index page and thread page. They are detected by page layouts characteristics.

Generic Crawler

In the generic crawler module, given a forum it enters into the thread page and it performs crawling where it avoids the duplicate links and page flipping links. It uses ITF algorithm and finds whether it is a valid thread page or invalid thread page.

Implementation

Implementation is the stage of the project(Phase- I) when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and its constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. The application is developed using Eclipse as frontend and SQL server is used as backend.

Figure 2: Screenshot for Login Page

Figure 2 depicts the login page. User has to enter the user name and password to enter into home page of forum.

Figure 3: Screenshot for Registration Page

Figure 3 shows the registration page. User has to enter the personal details to enter into login page.

Figure 4: Screenshot for Homepage of Forum

Figure 4 shows the homepage of forum. It contains the Mobile name, Threads, Posts, Viewers.

Figure 5: Screen shot for Posting the Query

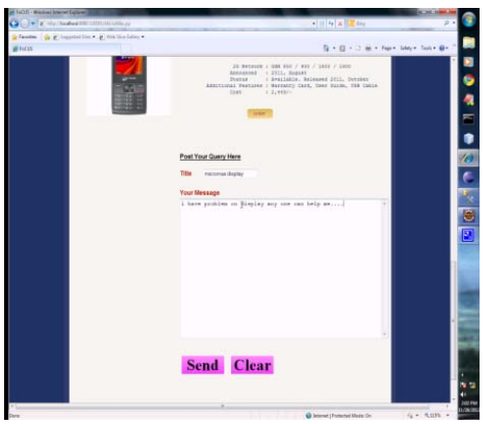


Figure 5 helps in posting the query.User enters into mobile forum and posts their query in title and message box.

Figure 6: Screenshot for Reply to the Query

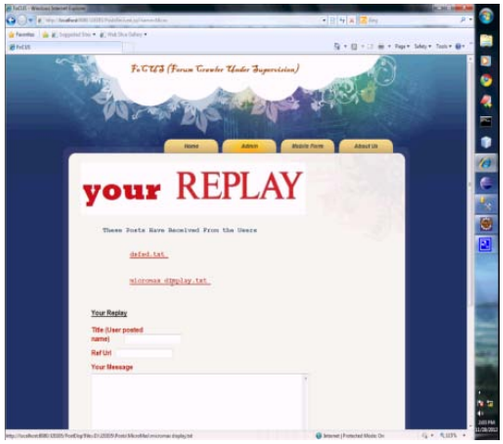


Figure 6 shows the screenshot for reply to the query. User can reply to the query to the user who has posted. It consists of title, ref URL, and your message.

Results and Conclusion

In the earlier sections, we discussed the requirement analysis, design, implementation and testing. This section discusses the result, conclusion and future enhancement.

Result

The objective of the project (Phase- I) "Web Forum Crawling using Index Thread Page Flipping Algorithm" is to check whether the URL link is valid or invalid. The index thread page flipping algorithm is used to find whether the link is valid or invalid. It consists of layout characteristics and regular expression patterns.

Conclusion

The URL recognition problem is identified and also the URL path of forums. Training set and regular expression patterns make the results applicable to many forum sites.

Future Enhancements

In future, we would like to discover new threads and refresh crawled threads in a timely manner and conduct more comprehensive experiments to further verify our approach and improve it.

References

Cai, R., Yang, J. M., Lai, W., Wang, Y., & Zhang, L. (2008). iRobot: An *intelligent crawler for web forums*. Proceedings of the 17th International Conference on World Wide Web (pp. 447-456).

Dasgupta, A., Kumar, R., & Sasturkar, A. (2008). *De-duping URLs via rewrite rules*. Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 186-194).

Manku, G. S. , Jain, A., & Sarma, A. D. (2009). *Detecting near duplicates for web crawling.*

Gao, C., Wang, L.,. Lin, C. Y., & Song, Y. I. Finding question-answer pairs from online forums. Proceedings of 31st Annual International ACM SIGIR Conference Research and Development in Information Retrieval (pp. 467-474).

Guo, Y., Li, K., Zhang, K., & Zhang, G. (2006). Board forum crawling: A web crawling method for web forum. Proceedings of 2006 IEEE/WIC/ACM International Conference on Web Intelligence (pp. 475-478).

Henzinger, M. (2006). Finding near-duplicate Web pages: a large-scale evaluation of algorithms. *Proceedings of 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 284-29)1.

Koppula, H. S., Leela, K. P. , Agarwal, A., Chitrapura, K. P., Garg, S., & Sasturkar, A. (2010). Learning URL patterns for webpage de-duplication. Proceedings of the 3rd ACM Conference on Web Search and Data Mining (pp. 381-390).

Schonfeld, U., & Shivakumar, N. (2009). Sitemaps: Above and Beyond the Crawl of Duty. Proceedings of 18th International Conference World Wide Web (pp. 991-1000).

Yang, J. M., Cai, R., Wang, Y., Zhu, J., Zhang, L., & Ma, W.Y. (2009). *Incorporating Site-Level Knowledge to Extract Structured data from web forums.* Proceeding of the 18th International Conference On World Wide Web, 181-190.

Zhang, L., Liu, B., Lim, S. H., & O'Brien-Strain, E. (2010). *Extracting and Ranking Product Features in Opinion Documents.*