System to Crawl Web and Forum for Medical Data

A.B. Shinde, A.S. Dange

Abstract: ***Medical Practitioners are those who require to have updated information of drugs, diseases, diagnosis and combinations of prescriptions. Internet is the rich source of information. The first attempt of every person is to search internet for his queries. But the relevant information searching and identifying the relevancy and accuracy of huge data is still a challenge. This paper focuses on data retrieval related to medical field. Forum crawler and web crawler will perform crawling of data from forums and web respectively. This system can be used by doctors as well as pharmacists to search medical related information as per their need.***

Keywords : Forum crawling, Web crawling, Effective Index

I. INTRODUCTION

Internet is a rich source of information. While searching any kind of information a user can get many links containing the information. It may happen that information relevant to users expectation may be scattered with different links. Proposed system will work for crawling of forums and information websites for medical practitioners, so they are able to search relevant information of drugs, diseases, diagnosis and combinations of prescriptions. Pharmacists also require information about drugs and to keep information about various drugs is an important task for them. So the proposed system will also be helpful for pharmacists. This system will contain a crawling manager which will be responsible for interaction with web and forum crawler.

II. LITERATURE REVIEW

A. Traversal strategy to direct the crawling of a given target forum

They have proposed a solution which identifies skeleton links and detect page-flipping links [2]. The main focus of skeleton links is on crawling valuable pages. It will also work to avoid pages which are duplicate or uninformative. The page-flipping links will focus on the process to download a long discussion thread.

B. Partial Tree Alignment based Data Extraction

This paper proposes a technique which will work for automatic web data extraction [3]. This method will align and extract data items from the records that have been previously identified. The only requirement is that the page should contain more than one data records.

C. Automatic Identification of Informative Sections of Websites

In this paper, four new algorithms are proposed namely:-ContentExtractor, FeatureExtractor, L-Extractor K-FeatureExtractor, [4]. These algorithms will work to identify primary content blocks. They will look for the blocks having desired features.

D. Extraction of structured data from web forum sites

This paper studied the problem of extraction of structured data from web forum sites [5]. For this, page level and site level knowledge was incorporated. Markov logic network was also used.

E. Concept of Anchor Points in retrieval of web documents

This paper introduce the concept of anchor points [6]. The main focus of this paper is to study the problem of anchor point indexing. For a user query anchor points could be a tiny set of key pages with the help of which the documents can be easily reached which are relevant to the user query.

F. Text Mining with Effective Pattern Discovery

In this paper, pattern discovery technique has been proposed [7]. This work refines the discovered patterns in text documents.

III. METHODOLOGY

A. System architecture

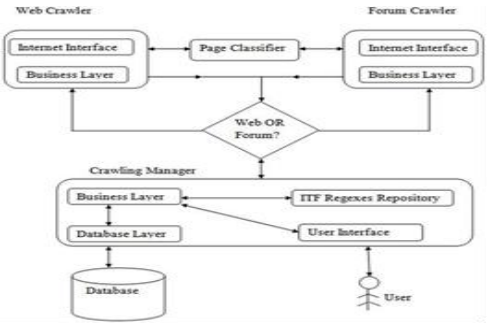

Fig 1. System architecture

Figure 1 shows the system architecture of proposed system. It consist of Forum crawler module, Web crawler module and Crawler manager module.

B. Modules

1) Crawling Manager Module:

Crawling manager module is the basic module of application. It will analyze, relate and store data from web and forum crawler. This module needs to handle different activities such as to manage the user interfacing of an application, to manage the topic entry and topic wise information repository, to check for new request, and take appropriate decisions about crawling, to manage the crawling calls with web and forum crawling modules, to support the analysis of data record

Revised Manuscript Received on January 15, 2020

Ms.A.B.Shinde, persuing Master of Engineering in Computer science at ADCET, Ashta.

Mr. A.S.Dange, Assistant Professor in Computer Science and Engineering Department at ADCET, Ashta.

2) Forum Crawler Module:

Forum crawler module will crawl data from forums. For this purpose it will perform following steps:-

1) Discovery of entry URL.
2) Identify and parse index page.
3) Implementing the related ITF Regxs for URL discovery.
4) Identifying the page flipping URLs.
5) Parsing of the Data out of Thread URLs.

3) Web Crawler Module :

This module will :-

1) Analyze the web information pages from entry URL.
2) parse the web information pages.
3) Retrieve the relevant information for input topic.

IV. EXPERIMENTAL RESULTS

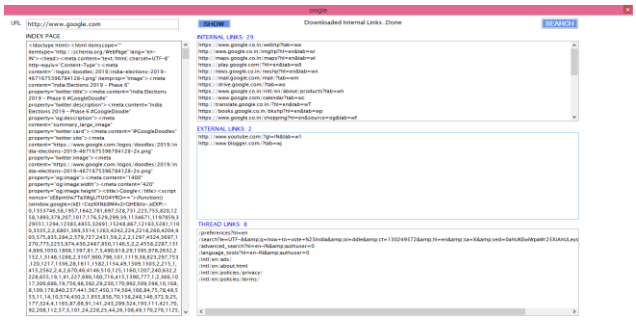

Fig 2. Downloading internal, external, thread links from a website

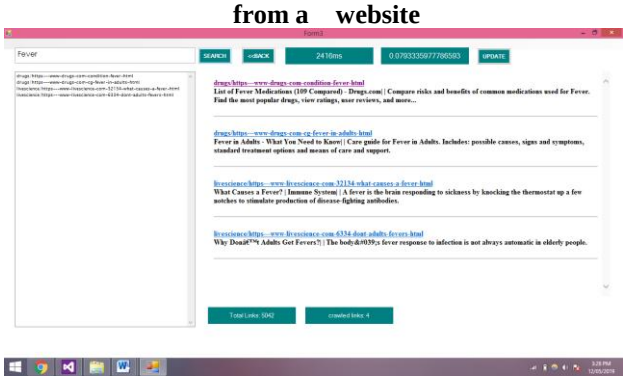

Fig 3. Searching information based on a keyword

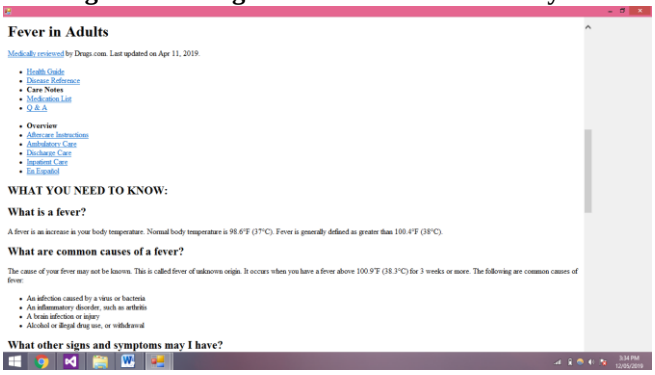

Fig 4. Search result

V. CONCLUSION

If a data retrieval technique is provided to users in online as well as offline mode, it will be effectively utilized. Proposed web and forum crawling system will help to a greater extent to medical practitioners and pharmacists also to search medical related information in both modes. A repository of information related to medical field will help to access the information to some extent in offline mode also. Thus system can be used in online and offline mode both.

REFERENCES

1. Jingtian Jiang, Xinying Song, Nenghai Yu, Member, IEEE, and Chin-Yew Lin, Member, IEEE, "FoCUS: Learning to Crawl Web Forums," IEEE Transactions On Knowledge And Data Engineering, Vol. 25, No. 6, June 2013.
2. 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. Intl ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 459-466, 2008.
3. 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.
4. Sandip Debnath, Prasenjit Mitra, Nirmal Pal, and C. Lee Giles , "Automatic Identification of Informative Sections of Web Pages," IEEE Transactions On Knowledge And Data Engineering, vol. 17, no. 9, September 2005.
5. 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 Intl Conf. World Wide Web, pp. 181-190, 2009
6. Ben Kao, Joseph Lee, Chi-Yuen Ng, and David Cheung, "Anchor Point Indexing in Web Document Retrieval, "IEEE Transactions On Systems, Man, And CyberneticsPart C: Applications And Reviews, Vol. 30, No. 3, August 2000
7. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012
8. Blog, http://en.wikipedia.org/wiki/Blog, 2012.

AUTHORS PROFILE

Ms.A.B.Shinde received Batchlor of Engineering degree in Information Technology from PVPIT, Budhgaon, Sangli in 2009 and persuing Master of Engineering in Computer science at ADCET, Ashta.

Mr. A.S.Dange received M.Tech degree in computer from BVUCOE, Pune in 2011 and persuing Ph.D in Computer Engineering from VTU, Belgaum. He is currently working as Assistant Professor in Computer Science and Engineering Department at ADCET, Ashta. His Ph.D research topic is Intelligent and Resource Redemption Crawling for World Wide Web. He is a member of ISTE and IEI.