An Efficient and Robust Metacrawler with Parallel Activities

Vimal Bibhu
Sr. Lecturer
Computer Science & Engg.,
Galgotia College of Engg. & Tech.
Greater Noida, U.P, India

Narendra Kumar
Sr. Lecturer
Computer Science & Engg.,
Galgotia College of Engg. & Tech. .
Greater Noida, U.P, India

Mohammad Islam
Sr. Lecturer
Computer Science & Engg.,
Galgotia College of Engg. & Tech
Greater Noida, U.P, India

Shashank Bhardwaj
Lecturer,
Master in Computer Application,
Krishna Institute of Engg. & Tech.,
Ghaziabad, U.P, India

***Abstract*— This paper presents the Metacrawler, a fielded Web service that represents the next level up in the information ``food chain.'' The Metacrawler provides a single, central interface for Web document searching. Upon receiving a query, the Metacrawler posts the query to multiple search services in parallel, collates the returned references, and loads those references to verify their existence and to ensure that they contain relevant information. The Metacrawler is sufficiently lightweight to reside on a user's machine, which facilitates customization, privacy, sophisticated filtering of references, and more. Standard Web search services, though useful, are far from ideal. There are over a dozen of different search services currently in existence, each with a unique interface and a database covering a different portion of the Web. As a result, users are forced to repeatedly try and retry their queries across different services. Furthermore, the services return many responses that are irrelevant, outdated, or unavailable, forcing the user to manually sift through the responses searching for useful information.**

Keywords- WWW: World Wide Web, HTML: Hypertext Markup Language, Hyperlink, Aggregation Engine, Parallel Web Interface ,

I. Background

Due to the variety of the structures and the sizes of today's web sites validating hyperlinks has become quite a difficult task" [1]. Since navigation throughout the website is done by usage of available hyperlinks, the quality of the web site can be reflected in the usage of its hyperlinks.

We are focusing on the link mining solutions for the WWW, specifically how it can be used for the hyperlinks evaluation. This lead to the terms such as link association which is defined as "rules that show the connectivity of different URLs" [2]. Another aspect of using link mining is using navigational link to extract navigational patterns, as Chen, Zaiane and Goebel define as "patterns discovered with web mining techniques" [3]. Navigational patterns can be used for different purposes, they can show how users of the web site behave in general or extract different (groups of) users' behaviors in order to adjust the web site to the need of a specific users group. The above pinpoints that link mining gives possibilities. However, the question is how these should be used. We will use the words knowledge discovery and data mining as synonyms, since in literature these names refer to the same techniques [4]. how these should be used. We will use the words knowledge discovery and data mining as synonyms, since in literature these names refer to the same techniques [5].

II. Problem discription

The World Wide Web has grown over the years to a size that became hard to foresee [6]. It is claimed that indexable WWW can have more than 11.5 billion pages [7]. It is no wonder that such a powerful tool attracts commerce and WWW gives: to the web users and to their owners is also the problem that the designer of the web site needs to face. Having many web sites to chose from, user can ditch one web site if it is to hard to browse [8], designers are challenged to fulfill the goal of users' navigational requirements. Beside the business needs for high quality, the web site quality evaluation is needed because of many existing ad-hoc solution for the web site design [9]. This continuous measure of quality is also required since, although designers are following standards of the high quality layout, still they can not predict the gap between their expectations and actual usage of designed site [10]. Support for the web site changes decisions is needed. Using the 1 outcome of the appropriate evaluation techniques one can improve existing site and if possible, use gained knowledge for future designs.

The problems that need to be solved are to identify where software engineers can find measures of the website and how to use them. Facing the fact that the users are the ones who evaluate the website, the designer should strive to validate design assumptions with the actual usage of the website. This type of assessment is only possible after "releasing" the website, since external quality of any software can be measured only starting with the moment when the software product is being used [11]. This leads to the problem of retrieving useful information from the usage logs and to make a relevant interpretation of it.

While discussing the hyperlinks usability issue one can consider the adequateness of the hyperlinks connectivity as the part of the hyperlink's utility value. How can we find the

answer to the question how to estimate utility of the hyperlinks using available data sources? Finally, if it is decided to validate quality attributes of the website, analytics need to decide which data need to be collected, use adequate methods for data processing and interpret the results of the process. Next to the problem how to interpret web site usage in order to estimate values of hyperlinks' utility is the question of reliability of used method. We are focusing on the hyperlinks used in the websites systems and the problem which consider evaluation of their utility. Since hyperlinks are supposed to reflect the relation between pages that they are linking, one can expect that there exist way to verify if users follow desired navigation paths. Although some tools can support the decisions for web site improvement based on the web site structure, the web designer still should understand how users are 'traveling' while using the web portal.

III. PREVIOUS EXISTING SYSTEM

The current system functions by accepting some inputs from user and then displaying result in the format not very intuitive. For E.g. In google.com after the searching the result they display the result in unformatted way where the user feel very inconvenient to navigate and proceed. In the current system in order to search a string user types same criteria at different sites until he/she gets the relevant links of given string. For Metacrawler user has also to type same keyword on different search engines to get the related links. The current system functions by accepting some inputs from user and then displaying result in the format not very intuitive. For E.g. In google.com after the searching the result they display the result in unformatted way where the user feel very inconvenient to navigate and proceed

In the current scenario whenever the user opens a job site like naukri.com, jobstreet.com etc, these sites provide some set of text boxes and Combo boxes to type or select searching criteria and then start searching But there are sites that provide to type criteria using comma or by using advance search options which a new user feels very uncomfortable.

Again for web search for a keyword there are chances that user won't get the links for which they are looking for, hence user move on to the next search engine.

.

IV. PROPOSED METACRAWLER

The main components are the User Interface, the Aggregation Engine, the Parallel Web Interface, and the Harness. The User Interface is simply the layer that translates user queries and options into the appropriate parameters. These are then sent to the Aggregation Engine, which is responsible for obtaining the initial references from each service, post-processing the references, eliminating duplicates, and collating and outputting the results back to the User Interface for proper display. Most of the sophisticated features of Metacrawler reside in this component. The Parallel Web Interface is responsible for downloading HTML pages from the Web, including sending queries and obtaining results from each search service [12]. The architecture of proposed efficient and robust metacrawler is given below.

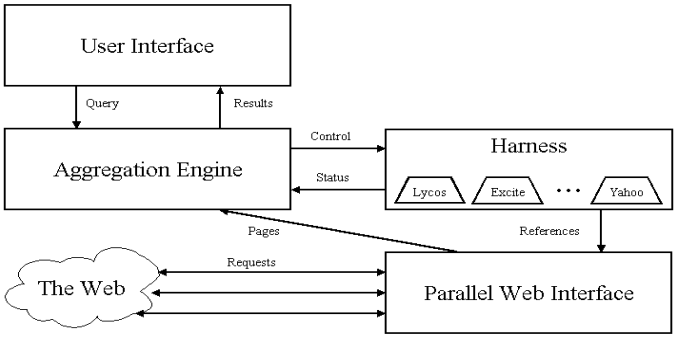

Figure 1. Architecture of Proposed Metacrawler

The Harness is the crux of the design; it's where the service-specific information is kept. The Harness receives control information detailing which references to obtain; it then formats the request properly and sends the reference to the Parallel Web Interface, which then returns a page to the Aggregation Engine. It also sends some status information back to the Aggregation Engine that is used to measure progress. The Harness is implemented as a collection of modules, where each module represents a particular service. It is designed so that modules can be added, modified, and removed without impacting the rest of Metacrawler.

V. FEATURES OF PROPOSED METACRAWLER

Metacrawler's architecture provides several advantages. It provides a layer of abstraction above traditional search services, which allows for greater adaptability, scalability, and portability as the Web grows and changes.

A. ADAPTIBILITY

Search services are extremely volatile in today's Internet. New search services are being launched continually. Almost as frequently is the rate at which search services upgrade their systems, which often means that their searching interface changes. Finally, while many search services are emerging, there are also a number of services being removed or moved to new sites, leaving and dangling references elsewhere in the Web. Metacrawler's modular design allows for new services to be added,

B. PORTABILITY

The Metacrawler is an intelligent interface to powerful remote services. It does not require large databases or large amounts of memory. This provides great flexibility in its location. Currently, Metacrawler exists as a universally accessible service at the University of Washington. However, we have created prototype implementations that reside on the user's

machine. We have also had great success in compiling Metacrawler on different architectures, including DEC OSF, Linux, and even Windows. The Metacrawler can run on most machines currently available with minimal effort, which means that most users can use Metacrawler, either locally or remotely, without needing to invest in expensive hardware. Further, Metacrawler is not rooted to any platform-specific technology, so as machines improve Metacrawler will be able to adapt and operate on those new machines as well.

C. SCALABILITY

As the Internet grows, so does its user base. It is important that the Metacrawler scale to handle these users. The load on the servers handling the Metacrawler service scales linearly with the number of requests per day. Unfortunately, given the expected usage projections of the Web, even linear scaling may require a daunting investment in hardware. However, using implementations resident on users' machines, we are able to scale without the need to add more machines. Every user is responsible for their own network and computational requirements, and they only need contact the server periodically to obtain updated service lists and other new information.

VI. PERFORMANCE ANALYSIS OF PROPOSED METACRAWLER

Performance of the proposed system based upon run time performance of software within the context of an integrated system. Performance is often coupled with stress testing and often require both hardware and software instrumentation. That is it is often necessary to measure resource utilization in an exacting fashion.

A. TIME BASED PERFORMANCE

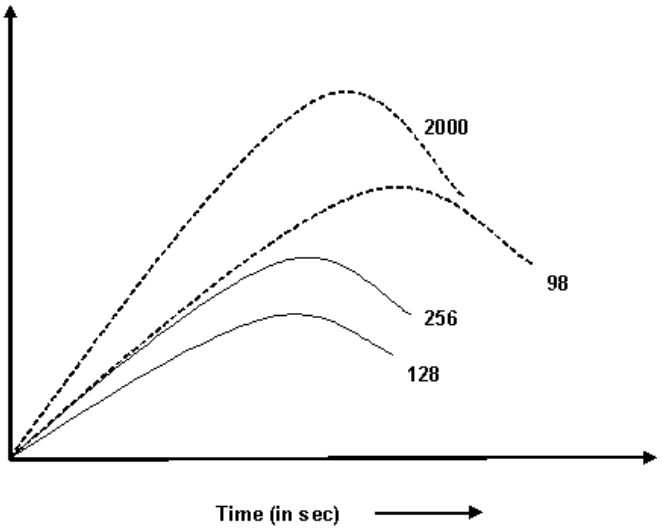

Figure 2. Performance (Time Based)

B. LOAD BASED PERFORMANCE

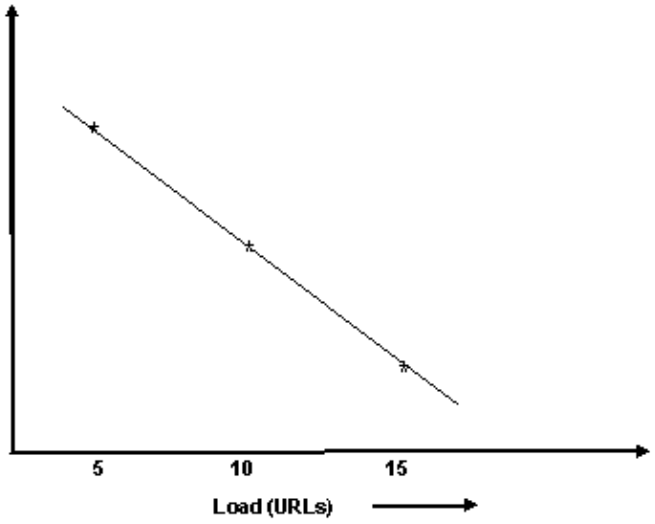

Figure 3. Performance (Load Based)

REFERENCES

[1] Erik Selberg and Oren Etzioni. MetaCrawler Home Page. URL: http:// www.cs.washington.edu/research/metacrawler.
[2] Erik Selberg and Oren Etzioni. Multi-Service Search and Comparison Using the MetaCrawler. In Proc. 4th World Wide Web Conference, Boston, MA USA, December 1995. URL: http: // metacrawler.cs. washington.edu: 8080/papers/www4/html/Overview.html.
[3] Mark A. Sheldon, Andrzej Duda, Ron Weiss, and David K. Gi_ord. Discover: A Resource Discovery System based on Content Routing. In Proc. 3rd World Wide Web Conference, Elsevier, North Holland, April 1995. URL: http://www psrg.lcs.mit.edu/ftpdir/papers/www95.ps.
[4] Erik Selberg and Oren Etzioni. MetaCrawler Home Page.
[5] Digital Equipment Corporation. AltaVista Home Page. URL: http://www. altavista.digital.com.
[6] Douglas R. Cutting, Jan 0. Pedersen, David Karger, and John W. Tukey. Scatter / Gather : A Cluster - based Approach to Browsing Large Document Collections. In Proceedings of the 1992 ACM SIGIR Conference, Copenhagen, Denmark, June 1992 URL: http://corp.excite.com/people/cutting/papers/sigir92.ps.
[7] Daniel Dreilinger. Integrating Heterogeneous WWW Search Engines. Master's thesis, Colorado State University, May 1995.
[8] Luis Gravano, Hector Garcia - Molina, and Anthony Tomasic. The E_ectiveness of GlOSS for the Text Database Discovery Problem. In Proceedings of the 1994 ACM SIGMOD Conference, Minneapolis, MN, May 1994.URL: ftp://db.stanford.edu/pub/gravano/1994/stan.cs.tn.93.002.sigmod94.ps.
[9] Inktomi, Inc. HotBot Home Page. URL : http://www.hotbot.com.
[10] Michael Mauldin. Lycos Home Page. URL : http://lycos.cs.cmu.edu.
[11] Michael L. Mauldin and John R. R. Leavitt. Web Agent Related Research at the Center for Machine Translation. In Proceedings of SIGNIDR V, McLean, Virginia, August 1994.
[12] Brian Pinkerton. Finding What People Want : Experiences with the WebCrawler. In Proc. 2nd World Wide Web Conference, Chicago, IL USA, October 1994.

AUTHORS PROFILE

Vimal Bibhu has received his Bachelor of Science in Chemistry(Hons.) from Magadh University, Bodh Gaya, Bihar, India, Post Graduate Diploma in Information Technology from IGNOU, New Delhi, India, Master in Computer Application from IGNOU, New Delhi, India and Master in

Technology in Computer Science & Engineering from CDAC Noida, Affiliated to Guru Gobind Singh Indraprastha University, New Delhi, India. He is working as Sr. Lecturer in the Department of Computer Science & Engineering at Galgotia's College of Engineering & Technology, Greater Noida, Uttar Pradesh India. He is a member of various Technical Societies viz. International Association of Computer Science & Information Technology (IACSIT), International Association of Engineers (IEANG). He published many research papers in various International Journals and Conferences.

Narendra Kumar has received his Bachelor of Technology and Master in Technology in Computer Science & Engineering from UP Technical University, Lucknow, India. He is working as Senior Lecturer in the Department of Computer Science & Engineering at Galgotia College of Science & Technology, Greater Noida, India. He is a member of various Technical Societies viz. Computer Society of India (CSI), Indian Society of Technical Education (ISTE). He published many research papers in various Conferences. His main research interests include: Wireless Sensor Network, Distributed & Mobile Computing and Middleware.

Mohammad Islam has received his Bachelor of Technology and he is also perusing Master of Technology in Computer Science & Engineering from Uttar Pradesh Technical University, Lucknow. He is working as Senior Lecturer in the Department of Computer Science & Engineering at Galgotia College of Engineering & Technology, Greater Noida, India. He is a member of various Technical Societies viz. Computer Society of India (CSI), Indian Society of Technical Education (ISTE). He published many research papers in various Conferences. His main research interests include: Wireless Sensor Network, Distributed & Mobile Computing and Middleware.

Shashank Bhardwaj has received his Bachelor of Technology in information Technology and he is also perusing Master of Technology in Computer Science & Engineering from Uttar Pradesh Technical University, Lucknow. He is currently working on the post of Lecturer in department of Master of Computer Application at Krishna Institute of Engineering & Technology, Ghaziabad, India. He is a member of various Technical Societies viz. Computer Society of India (CSI), Indian Society of Technical Education (ISTE). He published many research papers in various Conferences. His main research interests include: Wireless Sensor Network, Distributed & Mobile Computing and Middleware.