Designing a Regional Crawler for Distributed and Centralized Search Engines

MILAD SHOKOUHI
Department of Computer Engineering
Bu-Ali Sina University
Hamedan, IRAN

PIROOZ CHUBAK
Department of Computer Engineering
Sharif University of Technology
Tehran, IRAN

HASSAN BASHIRI
Department of Computer Engineering
Bu-Ali Sina University
Hamedan, IRAN

Abstract: - Today, by the growth of WWW, the significance and popularity of search engines are increasing day by day. However, today web crawlers are unable to update their huge search engine indexes concurrent to the growth in the information available on the web. Most of times they download some unimportant pages and ignore the pages that their probability of being searched is noticeable. This sometimes causes users to be unable to search in updated information. Regional Crawler that we introduce as new idea in this paper, improves the problem of updating and finding new pages to some extent by gathering users' common needs and interests in a certain domain, which can be as small as a LAN in a department of a university or as huge as a country. In this paper, we design the Regional Crawler architecture and introduce its application in centralized and distributed search engines

Key-Words: - Regional Crawler, Web Crawler Architecture, Multi-agent Systems

1 INTRODUCTION

1.1 The Significance of search engines

Surveys show that the use of internet is growing world wide both in servers and in clients. For instance, the number of web sites which was less than 3 millions in 1998 has become more than 9 millions by the year 2002[15]. The number of internet users which was less than 160 millions in 1998 increased to more than 600 millions by the year 2002[16 and 17].

Above statistics, show the huge amount of information on WWW and the growth of people's interests for this information in different parts of the world. The necessity of an efficient tool for finding what users are looking for makes the significance of search engines more obvious. Search engines take users' queries as input and produce a sequence of URLs that match the query according to a rank they calculate for each document on the web [12 and 13].

1.2 What are crawlers?

Web crawling is a process to collect the web pages that are interesting to search engine. it's usually a challenging task for general search engine [23] .Web crawler is a program that traverses the internet automatically by retrieving a web page and then recursively retrieving all linked pages [6 and 7].
It takes weeks to crawl the entire web because of huge amount of information on the it [5 and 8]. Even the largest search engines, like Google and Altavista, cover only limited parts of the web and much of their data are out of date several months of the year [19]. Therefore, they will not cover the information that changes hourly or daily (like news). Most of the recent works done on crawling strategies attempt to minimize the number of pages that need to be downloaded, or maximize the benefit obtained per downloaded page [19].

1.3 Regional Crawler Method

In this method, the crawling strategy is based on interests and user needs in certain domains. These needs and interests are determined according to common characteristics of the users like geographical location, age, membership and job. Regional Crawler uses these interests as the basic data for crawling strategy. The more a document has common interests of different domains, the more is its rank for being crawled. We have designed the architecture of regional crawler for two most common search engine structures called Centralized and Distributed.

2 A WEB CRAWLER DESIGN

The first crawler, Mathew Gray's Wanderer, was written in the spring of 1993, roughly coinciding with the first release of NCSA MOSAIC [8] and [9].

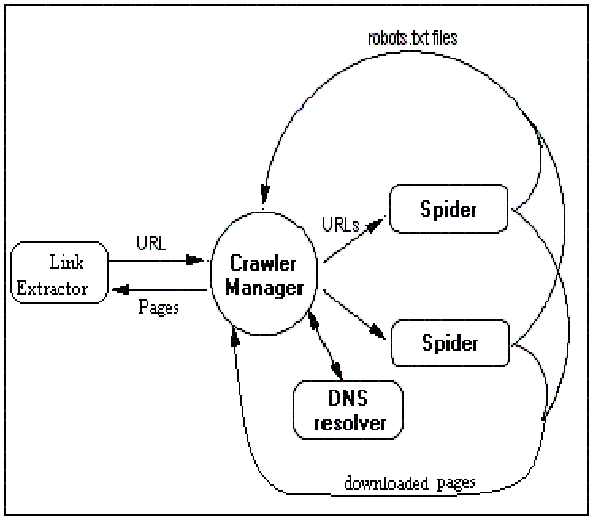

Fig. 1.Web Crawler Architecture

Figure 1 shows an easy architecture for web crawler:

- **Crawler Manager**: takes a set of URLs from Link Extractor and sends the Next URL to the DNS resolver to obtain its IP address. This saves a lot of time because spiders do not have to send requests to DNS every time they want to download a page.
- **Robots.txt file:** are the means by which web authors express their wish as to which pages they want the crawlers to avoid. Crawlers must respect authors' wishes as well.
- **Spider**: downloads robots.txt file and other pages that are requested by the crawler manager and permitted by web authors. The robots.txt files are sent to crawler manager for processing and extracting the URLs. The other downloaded files are sent to a central indexer.
- **Link Extractor**: processes the pages downloaded by the spider, extracts URLs from the links in those pages and sends the URLs to the crawler manager for downloading afterwards.

Any crawler must fulfill following two issues [19]:

1) It must have a good crawling strategy

2) It has to have a highly optimized system architecture that can download a large number of pages per seconds.

Most of search engines use more than one crawler and manage them in a distributed method. This has following benefits [18]:

- Increased resource utilization
- Effective distribution of crawling tasks with no bottle necks
- Configurability of the crawling tasks

3 USING REGIONAL CRAWLER IN CENTRALIZED SEARCH ENGINES

In centralized search engines [10], there is a central URL store, which sends URLs to the crawler for processing and download. The mechanism that leads to the production of a list of ranked URLs to get downloaded, determines the crawling strategy. There are three major crawling strategies for centralized crawlers in the literature [21].

- **Best-First crawlers**: use a queue in which the URLs are ranked according to their topic similarity and the pages there are found in.

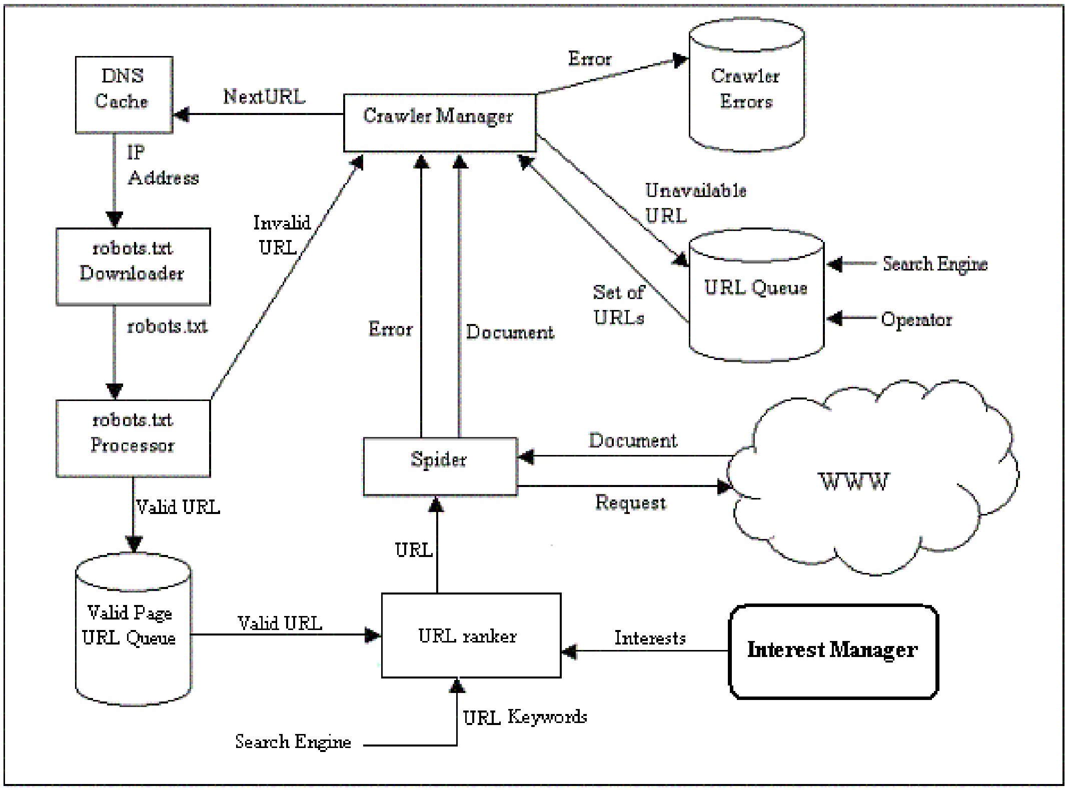

Fig. 2.Architecture of a Centralized Regional Crawler

- **Page ranking crawlers**: crawl the pages according to PageRank method described in [23 and 20]
- **InfoSpide**r: use neural networks to propagate the error and use the text near the links.

In this section, we will describe our crawling strategy, which is a new kind of Best-First strategy.

The main in detail architecture for crawlers in Centralized Search Engines is just like figure 2:

In this figure, valid URL Queue extracts the valid URLs from robot.txt and sends them to URL Ranker to rank the URLs according to user's interests for getting crawled more often. The Interest Manager has to specify the users' interests. A page processing unit in Search Engine extracts the keywords of pages and submits them to URL Ranker.

As we mentioned earlier the regional crawler decides which IP to download next according to the interests of the users in specific domains. Therefore, we must find a method to map users' interest to IP addresses that have sent their queries to the centralized search engine. The sequence of URLs that spider must download is determined by the common interests of a certain region. Each region is known by the IP addresses of the users with similar interests. These interests must be the specific characteristics of that domain; it means we ignore some public interests like Soccer, which is a common interest in most of regions. The regions granularity could be as small as a LAN or as big as a country. We map the IP addresses to the (interests of) domains according to the known characteristics of different environments such as country, location, age range, job and other common characteristics of the users in the domain.

As an example if the discussed domain is a CS department LAN in a university, the probability of requesting pages about computer science articles, other universities web sites and computer related topics is much more than a domain related to a bank. Therefore, we manually specify the interest of different regions in Region Interest table as illustrated in Table 1.

TABLE 1

REGION INTERESTS

Region	IP address	Region Interests
BASU CS LAN	163.218.216.1-100	Computer Engineering, Robocup, Soccer, etc.
Iran	Set of Iran IPs	Iran Politics, Wrestling

We obtain the keywords of each URL from the search engine. For example, http://ce.sharif.edu is the URL of computer engineering department of Sharif University of Technology in Iran and the retrieved keywords would be "Computer Engineering", "ACM Contest", and "Robocup" etc. Therefore, because of the similarity between region interests of Bu-Ali Sina University CS LAN(BASU LAN) and the discussed URL keywords, the URL Ranker should give a higher rank to this URL and its linked pages so that the BASU LAN spider, crawl it more often.

4 USING REGIONAL CRAWLER IN DISTRIBUTED AND AGENT BASED SEARCH ENGINES

4.1 Agent Based Search Engines

The current search engine architectures have a high cost [3] and some problems and limitations, which we will point to some of them in below: [4]

1. A central section has to calculate a large amount of information to answer the users' needs.

2. A bottleneck and a central error point will occur, so that if the central section faces an error, the whole system will crash!

3. According to the large amount of information and their complexity, the required processing is out of single central section abilities.

Nowadays to solve these problems, other kinds of search engines are designed and their improvement and popularity is increasing day by day.

Generally, a Multi-Agent architecture has five advantages over the past methods:

1) Flexibility
2) Robustness
3) Distributed Computing
4) Facilitating development/Maintenance
5) Facilitating Research

The Architecture of most Agent-Based search engines like [1, 2, 3 and 4] is based on a Three-Layer Model (Hermans, 1996). The main idea of this Three-Layer model is to divide the internet structure into three layers and devote some particular activities to each layer.

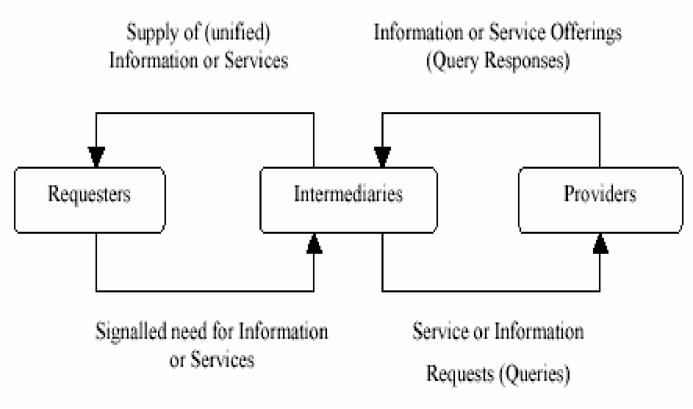
Supply of (unified) Information or Services
Information or Service Offerings (Query Responses)
Requesters
Intermediaries
Providers
Signalled need for Information or Services
Service or Information Requests (Queries)

Fig. 3.Three Layer Architecture

In this model (Figure 3), the requesters are the users who enter the query into the system and have an individual unique user profile. User profile contains the user interests and the results of previous searches [2]. Providers section also contains the services and information of the providers that are being searched for pages related to users' queries. Intermediaries are responsible for matching the users' requests with the information available from the providers or information which have become accessible by the other users according to their user profile.

4.2 Regional Crawler Agent Based Search Engines

As we explained before, current distributed and Agent-Based search engines are usually constructed based on a Three-Layer structure. Generally the main structure of a Personal Agent in most of Agent-Based search engines is just like figure 4[4].:

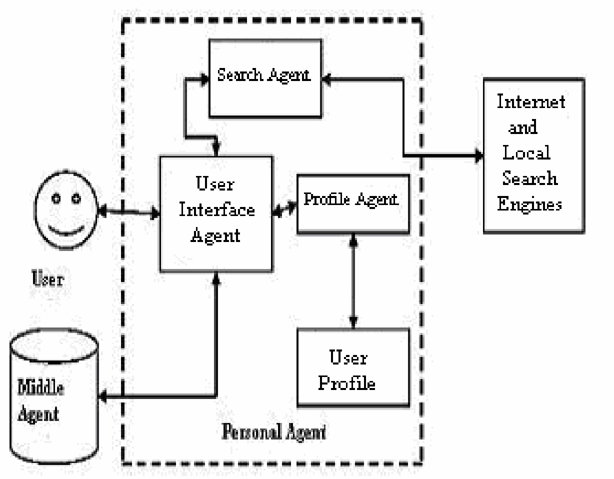

Fig. 4. Personal Agent in Agent-Based search engines

User Profiles are discussed in last section completely. Search Agent will search the internet and User Interface Agent acts as an interface between the user and the whole Personal Agent structure and enables the collaboration between the user and Agents for searching and entering the queries. Middle Agent plays the most important role in this architecture. It's a bridge between users and providers in the way that providers would announce the services that they provide and users will ask for their needs on the other hand and the Middle Agent would act as a Match Maker between those groups. The Advantage of this methodology is that some user profiles would be devoted to the users which show their needs, interests and past search results and when a query entered by a user and get ready for the search, Middle Agent would get the query and specify the subject of it, then it would search through the user profiles and User Agents to find a similar user according to the public profiles and get information from its past search results for the same or similar queries and send these results as an answer for the user who has entered the query. By adding the Regional Crawler as a Regional Agent to the above architecture we would have Figure 5:

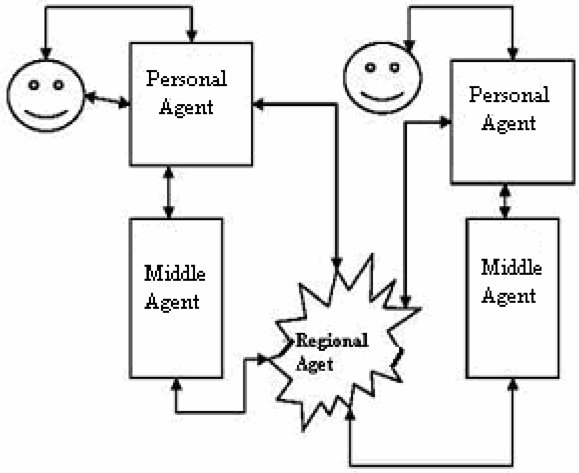

Fig. 5.Regional Crawler in Agent-Based Architecture

Regional Agent is responsible for collecting the users' interests of a specific region. Users with similar User Profile will be gathered together by Reinforcement Learning methods [14] or Supervised learning [11] depending on the Middle-Agent architecture. Regional Agent will search for users with similar interests and gather them in a unique public agent. Then we devote a special crawler for each regional agent and ask it to crawl the web in the way that it can satisfy the users' interests! This means that the crawler should look for the pages related to those regional interests before the other topics available on the web. Since the crawlers are in cooperation with Search Agents, Regional Agent will ask Search Agents to update the important web pages (look for important new pages) by announcing the user interests and needs to them. The important point in this architecture is that by implementing the RL methods, regions domain will be unlimited and as an example two Fans of a particular soccer club in two different locations of the world would be in the same region. So by adding a regional Agent to the above architectures we expect the important pages from the users (user agents) vision become updated more frequently. In this Agent-Based architecture, all the collaborations and cooperation and any massage exchanging between two agents is based on The Knowledge query and manipulation language KQML. KQML is a protocol for exchanging information and knowledge between agents [22]. According to this architecture the Regional Crawler shown as Regional Agent would be located in the Intermediaries Section discussed before.

5 FUTURE WORK

As the idea of a Regional Crawler is quite new in the literature there seems to be a lot of work to do on this topic. For example improving the profiling methods in agent-based crawlers by using reinforcement-learning [14] methods. In addition, realization of changes in a domain interests and adjusting the pages to be crawled is so important. For example in some weeks, people may get interested in political news in a domain according to the political situation occurring in the outside world relating to that domain, like a country.

Another idea is giving a higher rank to domains, which send more queries to our search engine. This causes more people to get better and more updated results than before.

6 CONCLUSION

In this Paper we designed a new web crawler called Regional Crawler and discussed about the role it plays in different search engine architectures. The main advantage of Regional Crawler over the other kinds of web crawlers is updating and finding the new most important pages according to users' needs and interests. So, in any specific Region the crawling method is affected by the results of retrieved interests.

As an example, the people of US are interested in baseball, so if any pages are added recently to a baseball site about a new tournament on next week, search engines should find these new pages instead of new pages related to the forecasting news of a small village in South Africa! The old methods don't cover this requirement, and all we have done is to improve the way of crawling to cover these requirements.

The more important a page is, the more frequently it will be looked for, updated and become available which current search engines are unable of giving such a service to users because of the lack in available pages about recently added pages in a specific host. If the search subject (query) be a member of that region's interests, the probability of getting answer for such a user query would increase.

References:

[1] T. Bauer, D. B. Leake, "Calvin: A Multi Agent Personal Information Retrieval System", *Department of Computer Science, Lindley Hall*, Indiana University, 2002.

[2] S. Pelletier, S. Pierre, H. Hoang, "Modeling a Multi-Agent System for Retrieving Information from Distributed Sources", *Journal of Computing and Information Technology-CIT* 11, 2003, 1, 1-10

[3] M. Chau, D. Zeng, H. Chen, M. Huang, D. Hendriawan, "Design and Evaluation of a Multi-agent Collaborative Web Mining System", *Department of Management Information Systems, Eller Collage of Business and Public Administration, The University of Arizona*, 2003

[4] B. Ling, "Enhanced Co-operative Knowledge Sharing Model for Agent Based Information Retrieval", *M.S.c Thesis, Staffordshire University*, 2000

[5] TGIF Google, http://www.stanford.edu/services/websearch/Google/TGIF/outof-index.html, Google at Stanford TGIF presentation 16-May-2003

[6] S. Mayocchi, "A Web Crawler for Automated Location of Genomic Sequences", *Department of Computer Systems and Electrical Engineering, University of Queensland, BA Thesis*, 2001

[7] P. Boldi, B. Codenotti, M. Santini, S. Vigna, "UniCrawler: A Scalable Fully Distributed Web Crawler", *in Proceedings of AusWeb02, the Eighth Australian World Wide Web Conference*, 2002

[8] A. Heydon, M. Najork, "Mercator: A Scalable, Extensible Web Crawler", *World Wide Web*, 2(4): 219-229, 1999

[9] Matthew Gray, Internet Growth and Statistics: Credits and Backgrounds, http://www.mit.edu/people/mkgray/net/background.html, 1996.

[10] P. Agars, "Architecture of a Search Engine", *An essay submitted to Dublin City University School of Computer Applications for the course CA437*: Multimedia Information Retrieval, 2002

[11] Chakrabarti S., "Data mining for hypertext: A tutorial survey, SIGKDD Exploration" *Newsletter of special Internet Group (SIG) on knowledge Discovery & Data Mining ,ACM* 1(2),pages 1-11,2000

[12] M. Hollander, "Google's Page Rank Algorithm to Better Internet Searching", University of Minnesota, Morris Computer Science Seminar, Spring 2003

[13] G. Salton, M.J. McGill, "Introduction to Modern Information Retrieval", *McGraw-Hill Book Co., New York*, 1983

[14] Sutton R. S., Barto A. G., "Reinforcement Learning: An Introduction*", MIT Press, Cambridge, MA*, 1998

[15] Web Statistics (size and growth), http://wcp.oclc.org/stats.html, 2002.

[16] Statistics – Web Growth, http://www.upsdell.com/BrowserNews/stat-growth.html, 2003.

[17] Internet Domain Survey, http://www.isc.org/ds/WWW-200301/index.html, Jan 2003.

[18] A. Jain, A.Singh, L. Lin, "A Scalable, Distributed Web-Crawler"

[19] T. Suel, V. Shkapenyuk, "Design and Implementation of a High-Performance Distributed Web Crawler", 2002

[20] Brin S. Motvani R., Page L. and Winograd T., "What Can You Do with the Web in Your Pocket", *Bulletin of IEEE Computer Society Technical Committee on Data Engineering*, 1998

[21] Menczer F., Pant G., Srinivasan P. and Ruiz M., "Evaluating Topic-Driven Web Crawlers", *In Proceedings of the 24th annual International ACM/SIGIR Conference, New Orleans*, USA, 2001

[22] M. N. Huhnus, Larry M. Stephens, "Multi Agent Systems and Societies of Agents, Multi Agent Systems A Modern Approach to Distributed Artificial Intelligence", *edited by Gerhard Weiss ,The MIT Press Cambridge,*

Massachusetts London, England, 2nd printing , 2000

[23] S. Brin, L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Science Department, Stanford University, Stanford, USA*, 1998