Design of a Least Cost (LC) Vertical Search Engine based on Domain Specific Hidden Web Crawler

Sudhakar Ranjan, Department of Computer Science Engineering, Apeejay Stya University, Gurgaon, India
Komal Kumar Bhatia, Department of Computer Engineering, YMCA University of Science & Technology, Faridabad, India

ABSTRACT

Now days with the advent of internet technologies and ecommerce the need for smart search engine for human life is rising. The traditional search engines are not intelligent as well as smart and thus lead to the rise in searching costs. In this paper, architecture of a vertical search engine based on the domain specific hidden web crawler is proposed. To make a least cost vertical search engine improvement in the following techniques like: searching, indexing, ranking, transaction and query interface are suggested. The domain term analyzer filters the useless information to the maximum extent and finally provides the users with high precision information. Through the experimental result it is shown that the system works on accelerating the access, computation, storage, communication time, increased efficiency and work professionally.

KEYWORDS

Domain Specific Hidden Web Crawler (DSHWC), Domain Term Analyzer, Frequency Calculator, Indexing, Repository, Web Log

1. INTRODUCTION

The hypertext was introduced in July of 1945 after around 50 years web sites began. As huge amount of information is available on Web and the numbers of web sites are increasing, the quantity of pages are also increasing more rapidly. The information stored over the web is accessible through the internet. Web pages over WWW (World Wide Web) are generally classified into static and dynamic pages. The static / fixed pages fall under category of surface web and dynamic pages fall under hidden web category. As the volume of hidden web is growing exponentially, a lot of time is spending by the user in searching relevant web pages. For accessing, searching and retrieval of the web information a search engine is generally required. The conventional search engines classify and index only static pages. The dynamic pages are not indexed by the conventional search engines. The general purpose search engine does not work effectively for finding the topic of relevant search over the hidden web. This drawback has been removed after the development of the vertical search engine which operates on the principle of finding the topic relevant pages there by leading to the better quality of web search for Hidden Web. The “Hidden Web” or “Invisible Web” contents are currently not a part of conventional search engine. The searching of the hidden web is very difficult due to the two basic
reasons. The first issue is size of the content stored in online database and secondly it requires access to the database through restricted search interface so as to extract relevant content. This increases the cost of accessing, searching and retrieval. Hence, there is a need to design and develop a Least Cost Vertical Search Engine for hidden web that can reduce the cost of crawling, accessing and storing along with communication cost for searching of the hidden web contents.

2. RELATED WORK

In Shettar and Bhuptani (2008) the vertical search engine based on domain classifier is built on seven modules: crawler (spider), HTML parse, filter, domain classifier, page ranker, URL db, search interface. In Peshave (2005) the work on structured-data on the web has focused mostly on providing users access to the data. However, the significant value can be obtained from analyzing collections of meta-data on the Web. Desa (2007) describes in detail the basic tasks a search engine performs. An overview of how the whole system of a search engine works is provided. A WebCrawler application is implemented using Java programming language. In Raghavan and Garcia-Molina (2001) a large amount of on-line information resides on the invisible web – web pages generated dynamically from databases and other data sources hidden from current crawlers which retrieve content only from the publicly Indexable Web. Specially, they ignore the tremendous amount of high quality content “hidden” behind search forms, and pages that require authorization or prior registration in large searchable electronic databases.

Zhou, Xiao, Lin, and Zhang (2010) provides a framework for addressing the problem of extracting content from this hidden Web, built a task-specific hidden Web crawler called the Hidden Web Exposer (HiWE), describes the architecture of HiWE. In Huitema and Fizzano (2010) a distributed template-customized vertical crawler which is specially used for crawling Internet forums. The performance of centralized vertical crawler and distributed vertical crawler are compared in the experiment. In Li, Zhaol, and Huang (2010) the focus is on the tasks of crawling and indexing a large amount of highly relevant Web pages, improved HITS algorithm combining link value with topic similarity highlights the difference of links and it assigns different weights to different links. In Bhatia and Sharma (2008), The framework extracts hidden web pages by accrue benefits of its three unique features: 1) automatic downloading of search interfaces to crawl hidden web databases, 2) identification of semantic mappings between search interface elements by using a novel approach called DSIM (Domain-specific Interface Mapper), and 3) the capability to automatic filling of search interfaces. The effectiveness of proposed framework has been evaluated through experiments using real web sites and encouraging preliminary results were obtained.

3. SYSTEM ARCHITECTURE

In this section the proposed architecture of vertical search engine based on domain specific hidden web crawler is provided. The major functions are: Hidden web crawler, Repository, Indexer, Index, Result Merging Rank Calculator, Query Interface, Query Processor, Query Term Generator, Domain Identifier Analyzer, Term Matcher, Query Log, Web Log, Frequency Calculator and Next Query Predictor. In design of a least cost (LC) vertical search engine based on domain specific hidden web crawler through parallel computing reduced the cost. Out of four cost component, the most important is the access cost, storage cost and communication cost. The cost minimization depends on the size and type of application. In the parallel system, the communication cost is minimizing as because many sites are involved for the data transfer. If the data can be completely stored in main memory, then computation cost is reduced. Hidden web crawler and Index Integrator load balancer distributes the hardware and software traffic load. Once the hardware and software load is distributed then systems work efficiently and effectively and also running cost of the access time, memory time, communication time, storage time and computational time is reduced. In indexing segmentation and
The Impact of the Mode of Data Representation for the Result Quality of the Detection and Filtering of Spam
