A Parametric Layered Approach to Perform Web Page Ranking

Ratika Goel
Dept. of Computer Science
Amity School of Engineering and Technology
Amity University, India

Anchal Garg
Dept. of Computer Science
Amity School of Engineering and Technology
Amity University, India

ABSTRACT
Web crawling is the foremost step to perform the effective and efficient web content search so that the user will get the specific web pages initially in an indexed form. Web crawling is not only used for searching a webpage over the web but also to order them according to user interest. There are number of available search engines and the crawlers that accept the user query and provide the page search. But, there is still the requirement and scope to improve the search mechanism. In this present work, dynamic and user interest evolution based parametric approach is defined to perform the web crawling and to arrange the web pages in more definite way. In this work a layered approach is defined, in which the initial indexing will be performed based on the keyword oriented content match and later on the indexing will be modified based on user recommendation. The presented work will provide an recommendation based web page indexing so that effective web crawling will be performed.

Keywords
Crawling, Indexing, Recommender system

1. INTRODUCTION
A web crawler is the heart of search engine that work actively as the central part of the search engine. A crawler actually performs the web search in a fraction to perform the related content search. The efficiency and the reliability of a search engine actually depend on the efficiency vector of a web crawler. A search engine actually passes the user query to the web crawler and the crawler search the information over the public webpages. The work of web crawler is to process this query and identify the keywords from the query. By using these keywords the page search will be performed over the web. As we can see in figure 1, the basic architecture of the web crawler is presented. The crawler used a queue and a scheduler to perform the dynamic web search evaluation and fix storage for the statistical analysis.

There are number of algorithmic approaches provided by different authors to process on user web request. The main challenges faced by any algorithmic approach is given as under

A) Scale
The web is very huge and frequently evolving. Crawlers that search for broad coverage and good newness must achieve extremely high throughput, which poses many difficult engineering problems. Present search engines running the multiple servers simultaneously to handle maximum number of user request without the service delay.

B) Content Specific Issues

It is not possible for any crawler to perform content check on

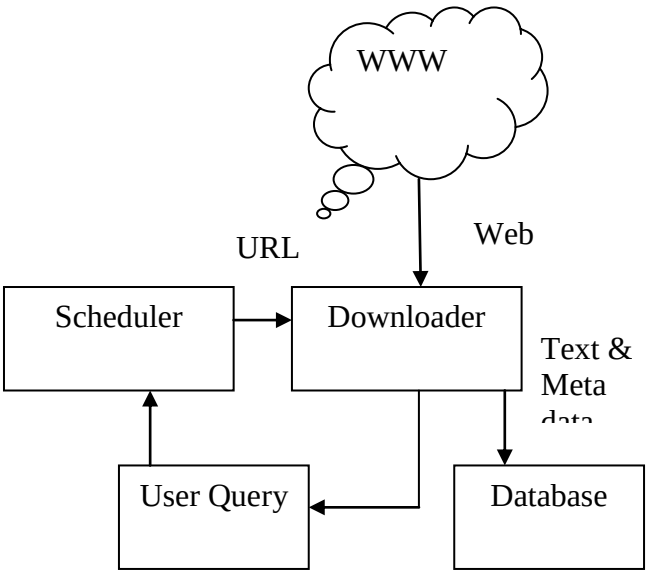

Figure 1: Web crawler architecture

all sites over the web. The another criticality is the updation of the web contents regularly. Because of this the selective crawling is required. Because of this content based search is performed over the web.

C) *Social Obligations*

Crawlers does not added much extra load on a website while performing the crawling. They uses the safety mechanism so that the high throughput from the crawler will be obtained.

D) Adversaries
The search engines also filter the web contents before presenting them. The filtration is performed respective the duplicate contents and the relevancy to the user requirements.

1.1 DATA DETECTION APPROACHES

The main objective of the web crawler is to identify the data over the web based on user query filtration as well as user query analysis. Some of such approaches are discussed in this section

A) Copy Detection Approach
In such approach, the documents are being searched over the web based on the perfect match. These kinds of approaches are being used by many plagiarism detection software systems. They basically perform the data check based on the different coping detection forms such as sentence based check, word based check, paragraph based check etc. Some of the crawler also performs the word substitution as well as the partial

sentence content extraction to perform the data match. These kind of systems also include the Web documents and evaluating its performance to detect similar documents, sentence-based copy detection is an approach which is applied to determine whether sentences in any two documents should be treated as the same or different according to the degrees of similarity of the sentences, which is computed by using the three least frequent 4-gram approach.

B) Name Entity Based Recognition

The term Named Entity is extensively used in Natural Language Processing. Named Entity Recognition is developed as a subtask of Information Extraction, because information units like names including person, location and organization names, and numeric expressions including time, date, money and percent expressions are the key points for information extraction. Extraction of named entities from text is simple for humans. People firstly use orthographic rules in order to find named entities by looking at the first letter of a word. If it starts with a capital letter, then it is a candidate for a named entity. Up to this point this process is also simple for computer, but it will identify a word as a named entity if it starts with a capital letter and in fact it is not a named entity. At this point, contextual clues are used to recognize named entities which they do not met before. For named entity recognition, there are two approaches from the point of view of computer: rule based approach and machine learning approach.

C) Rule Based Algorithm

In rule-based approach, the entities are analyzed by experienced linguistics and hand-crafted rules are created. In order to extract entities mainly three phases are used: Linguistic Preprocessing, Named Entity Identification and Named Entity Classification. Linguistic Preprocessing includes tokenizing, part of speech tagging, stemming and using the list of known names (database lookup). In order to identify named entities, boundaries of each named entity are detected. This includes the start and end structure of all the words that can be thought as named entity. In this possible named entities are generated by using punctuation marks or capitalization. Also, entities consisting of more than one word are identified at this stage. When possible named entities are identified, classification begins. Classification is performed in three stages: application of rules, database lookup classification and considering the matching of classified named entities with the unclassified ones. Rules are generated by experienced linguists. Rules are formed considering appositives or certain keywords that can precede or succeed a possible name. Classification starts by matching possible named entity with the generated rules. If there is no match with the rules, then database lookup is used. In these two stages, system's aim is to define exact category of a named entity. If classification cannot be performed in the previous two stages, then partial matching strategy is used as a final stage. This stage tries to identify truncated forms of names.

D) Machine Learning Based Approach

Machine learning approach is performed mainly in two stages: feature extraction and feature selection. In the feature extraction stage, previously generated training corpus is used. In this training corpus names and their categories are previously labeled. By using training corpus, features are extracted and classifier is trained with examples of sample names and their categories. After the classifier is trained by using training corpus, the system at this stage is tested by the real input. This time system tries to identify the category of unseen data. Machine learning approaches can be separated into three categories as supervised learning (SL), semi-supervised learning (SSL) and unsupervised learning (UL).

E) Supervised Learning

In SL the main purpose is to teach the system features of positive and negative examples on a large collection of annotated documents. SL is the most common approach used in NER for machine learning approach. For this purpose specific machine learning algorithms are used: Hidden Markov Models (HMM), Maximum Entropy Models (ME), Decision Trees, Support Vector Machines (SVM) and Conditional Random Fields (CRF). HMM tries to predict hidden parameters from observable parameters. All these techniques are used in systems that read a large annotated training collection and create disambiguation rules. These rules are then applied to a different test collection to identify named entities.

F) Semi-Supervised Learning

SL needs a large annotated corpus and it is not always possible to create such a corpus and preparing that kind of corpus is a very time consuming task. For this reason researchers prefer another option to perform named entity recognition work and this option is Semi-supervised Learning. Semi-supervised can also be called as weakly supervised and main technique for this approach is bootstrapping. In bootstrapping a small number of examples are given to the system and then system tries to find related sentences and contextual clues with the given examples. This process is iteratively applied in order to make the system find new clues with the help of newly discovered examples.

G) Technique Based on Similarity Measures

Techniques using similarity measures calculate a similarity value for each document pair and in order to understand a document is similar to another one its similarity value has to exceed some threshold value. In approaches using similarity measures the value associated with threshold is very important. Specifying a small value for the threshold will bring on false alarms in the case of plagiarism detection and unrelated documents will be identified as plagiarisms. On the contrary specifying a high value for the threshold will cause documents that are really plagiarisms to be missed. Several efforts have been made by researchers for determining the similarity of a document to another document.

2. EXISTING WORK

Avanish K. Singh[1] focused the work on the web crawler performance that is one of the major web issue. Author defined an approach to obtain higher throughput even in overloading conditions. He has improved the growth rate of automation of the work so that the optimal solution will be drawn from the system. Another kind of work done over the web page crawling is with the inclusion of intermediate agents. These agent based work is defined by P. Srinivasan[2]. An agent is the inter-mediator between the web server and the client. The research done by the author focuses to improve the ability to access the web information. The author worked on a focused crawler to retrieve the information in the area of biomedical information whose relevance is assessed using both genetic and ontological expertise. The agents defined in the work will not only perform the information retrieval but also perform the analysis on the process of information fetching. Author has also discussed the different issues and challenges related to the agent based retrieval from the system. To access the web information effectively as well as to identify the duplicate pages over the web, a smart work is defined by J. Cho[3]. The author has defined the web crawling and web searching under the criteria of replicated data and the web pages. The author defined an approach to identify the rank based analysis for the duplicate data over the web. Author has defined an efficient approach to identify these duplicate web pages and the hyperlink over the

millions of web pages present over the web. Author also presented a new work to optimize the solution by using the multiple crawlers simultaneously called parallel Crawlers[4]. The author has defined how these web documents are collected and also resolve the challenge of replicated data search over a huge amount of web pages. Author has defined a multiple architecture to perform the parallel search over the web and to combine and filter the search results. Author work is divided in two main phases, first to perform the parallel crawl search and retrieve the results and then to perform a metric based analysis over the results to retrieve the meaningful information and to avoid the duplicate search.

O. Brandman has explored the relation between the web server and the crawler. The work is focused on the regular search performed by the users over the web as well as focused on the performance metrics collected to analyze the outcome. Based on these performance vectors, the author also drive the approach that utilizes the bandwidth effectively and to perform a keyword based match over the web. Author has defined a filtered search over the web based on the meta data analysis of a webpage. Author proposes that web servers export meta-data archives describing their content[5]. Charu C. Aggarwal presented a focused crawler work based on topic based search over the web. Author has discussed recent techniques to obtain effective web crawling for the specific topics. Author has mainly focused the work on two major web crawling approach called the Intelligent Crawling Methods and User Centric methods. Intelligent Crawling Methods is defined by a pattern based search based on the statistical analysis to derive the linkage information. The user based analysis is performed to analyze the user search analysis and the user recommendation to perform the effective and ranked crawling over the web[6]. Author has also discussed some more recent algorithms to perform topic based search effectively over the web.

V. Shkapenyuk describes the design and implementation of an effective distributed web crawler so that the work will be implemented on a private network. In such case the major issue is the fastest and the most relevant search over the workstation along with workstation specification. Author presented an architecture for the system with the performance bottleneck and to drive the high performance based association search over the web[7].The author has defined the work under the capabilities of the web application layer and suggest some modification so that the rule based search will return more effective results from the system[8] Hussein Issa (2010) studied the problem of duplicate web contents and define the interest in the business world in the form of duplicate payments etc. The author has discussed different such cases where the fraud is done because of duplicate payments so that huge amount of money is lost[9]. Hani Khoshdel Nikkhoo (2010) stated that near-duplicate documents can adversely affect the efficiency and effectiveness of search engines. Due to the pair wise nature of the comparisons required for near-duplicate detection, this process is extremely costly in terms of the time and processing power it requires. Despite the ubiquitous presence of near-duplicate detection algorithms in commercial search engines, their application and impact in research environments is not fully explored. The implementation of near-duplicate detection algorithms forces trade-offs between efficiency and effectiveness, entailing careful testing and measurement to ensure acceptable performance and described a scalable implementation of a near-duplicate detection algorithm, based on standard shingling techniques, running under a Map Reduce framework. Also explored two different shingle sampling techniques and analyze their impact on the near-duplicate document detection process. In addition, investigated the prevalence of near-duplicate documents in the runs submitted to the adhoc task of TREC 2009 web track [10]. J Prasanna Kumar and P Govindarajulu (2009) reviewed that the development of Internet has resulted in the flooding of numerous copies of web documents in the search results making them futilely relevant to the users thereby creating a serious problem for internet search engines. The outcome of perpetual growth of Web and e-commerce has led to the increase in demand of new Web sites and Web applications. The survey paper intended to present an up-to-date review of the existing literature in duplicate and near duplicate detection of general documents and web documents in web crawling. A brief introduction of web mining, web crawling, and duplicate document detection is also presented [11]. Erkan Uyar et al (2009) proposed a new near-duplicate news detection algorithm: Tweezer. In this algorithm, named entities and the words that appear before and after them are used to create document signatures. Documents sharing the same signatures are considered as a near-duplicate. For named entity detection, introduced a method called Turkish Named Entity Recognizer, TuNER. For the evaluation of Tweezer, a document collection is created using news articles obtained from Bilkent News Portal. In the experiments, Tweezer is compared with I-Match, which is a state-of-the-art near-duplicate detection algorithm that creates document signatures using Inverse Document Frequency, IDF, values of terms[12].

3. PROPOSED WORK

The effectiveness of a search engine system is based on the degree of user satisfiability respective to the obtained search results. The presented approach is in the same direction to obtain the results more user oriented. To get such kind of improvement, we have defined a user feedback system as well as the user interest analysis order the search results. To present the results effectively, a user oriented indexing is performed. The presented work has improved the search engine system in two ways.

(i) An improved Filtered Keyword based crawling approach

(ii) A User Based Ranking System

The actual work of the processing stage depends on the crawling. The crawling is here performed upto two levels i.e. to identify the inner links of a web page. Here figure 2 is showing the basic processed follow by the web crawler.

According to this process, at first the url links will be identified based on the user query. Now for each link the web server will be intimated to extract the web contents. Here the server robot will check the user query and the user validity. If all are satisfactory then it will allow to access the page. As the page is fetched from the url, it is processed again to extract the web links. These web links represents the inner links of a web page and processed in same way. Finally as the links are verified the extraction of the text will be performed and relatively the index list will be maintained and links list will be updated.

In the first phase, as the user pass a search query, a filtration over the query is performed to identify the keywords only. These keywords are further analyzed in terms of prioritization, keyword type and the frequency. Based on these vectors the pruning is performed to avoid the rarely used keywords. Once the Keyword categorization is done, the prioritization is performed based on the keyword type. Now this keyword based query is used to perform the effective web search. The algorithmic flow of the approach is given as

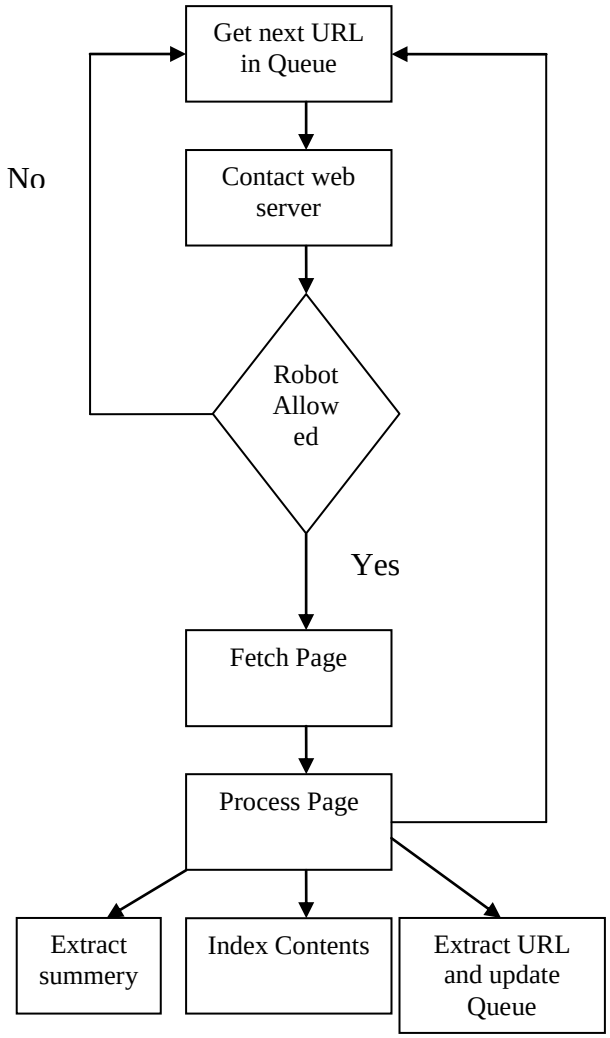

Figure 2 : Basic Web Crawling Process

Algorithm(UserQuery)
{

(i) Convert the Userquery to the WordList

(ii) Perform the keyword extraction by removing the stop list words

Keywords= WordList – StopList

(iii) Categorize the keywords in different types like Abbreviation, Date, Number, Text

(iv) Assign the priorities to different type of keywords

(v) Identify the frequency of different kind of keywords

(vi) Find Average Keyword Frequency

(vii) Prune the keyword list by eliminating the rarely used keywords

(viii) Use the obtained list as the search query.

}

The presented approach not only improves the search mechanism but also increase the length of keywords passing to the search system. As the search will be performed based on these keywords, in the second phase the work is to present these keywords in an index form. The indexing of the search query will be based on two main vectors

(i) User Feedback

(ii) User interest Analysis

The user feedback will be taken in a database to identify the priority of the pages that user want to see in earlier pages. The interest analysis is the intelligent system, in which user page visit history will be used to identify the user interest.

Based on these two vectors the page ranking will be performed. As the work includes both the analytical and the user interest vectors so it is assumed that the approach will return the effective results under the user search criteria.

4. CONCLUSION

The presented work is the improvement of existing crawling and the ranking process by implementing a layered scheme. In which first layer will do an analytical task to filter the query in effective way so that more accurate results will be driven from the system and in second stage the prioritization approach is defined to present the result in effective way.

6. REFERENCES

[1] Avanish K. Singh,"Novel Architecture of Web Crawler for URL Distribution", International Journal of Computer Science and Technology, Vol 2, Issue 3, Sept 2011, pp 42-45

[2] P.Srinivasan,"Web Crawling Agents for Retrieving Biomedical Information", ACM, NETTAB 2002 Bologna, Italy, pp 1-8

[3] J. Cho,"Finding replicated web collections", In proceedings of the 2000 ACM international conference of Management of Data (SIGMOD) 2000 pp 355-366

[4] J. Cho," Parallel Crawlers", In proceedings of WWW2002, Honolulu, hawaii, USA, May 7-11, 2002. ACM 1-58113-449-5/02/005

[5] O. Brandman, "Crawler-Friendly Web Servers", In Workshop on Performance and Architecture of Web Servers (PAWS), June 2000, pp 1-16

[6] Charu C. Aggarwal,"On Learning Strategies for Topic Specific Web Crawling", WI '06 Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence

[7] V. Shkapenyuk,"Design and Implementation of a High-Performance Distributed Web Crawler", In Proceedings of the 18th International Conference on Data Engineering (ICDE 2002), pp. 357–368

[8] B. Polverini,"Levels of Awareness: Design Considerations for Web Crawlers and Censorware Detection", White paper, Princeton University, May 2011

[9] Hussein Issa Rutgers Business School, Rutgers University "Application of Duplicate Records detection Techniques to Duplicate Payments in a Real Business Environment"

[10] Hani Khoshdel Nikkhoo "The Impact of Near Duplicate Documents on Information Retrieval Evaluation" by A thesis presented to the University of Waterloo in fulfillment of the thesis requirement for the degree of

Master of Mathematics in Computer Science Waterloo, Ontario, Canada, 2010

[11] J Prasanna KumarProfessor, "Duplicate and Near Duplicate Documents Detection: A Review". European Journal of Scientific Research ISSN 1450-216X Vol.32 No.4 (2009), pp.514-527

[12] Cho, Junghoo; Hector Garcia-Molina (2000). "Synchronizing a database to improve freshness". Proceedings of the 2000 ACM SIGMOD international conference on Management of data. Dallas, Texas, UnitedStates: ACM. pp. 117, 2009