ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

Improving the Efficiency of Web Crawler by Integrating Pre- Query Approach

Vishakha Shukla

M.E, Dept. of Computer Science and Engineering, Rungta College of Engineering and Technology, Bhilai (C.G.), India

ABSTRACT*:* As deep web grows at a very fast pace, there has been increased interest in techniques that help efficiently locate deep-web interfaces. However, due to the large volume of web resources and the dynamic nature of deep web, achieving wide coverage and high efficiency is a challenging issue. Proposed approach focuses on improving the efficiency of web crawler by integration of pre-query processing approach in order to leave irrelevant query thereby enabling crawler to take out most relevant links in fast and efficient manner. Also using post-query approach Crawler achieves fast in-site searching by excavating most relevant links with an adaptive link-ranking.

KEYWORDS: Crawler, Web Crawling, World Wide Web, Depth First Search, Spider

I. INTRODUCTION

Phishing To find information on the hundreds of millions of Web pages that exist, a search engine employs special software robots, called spiders, to build lists of the words found on Web sites. When a spider is building its lists, the process is called Web crawling. In order to build and maintain a useful list of words, a search engine's spiders have to look at a lot of pages. The importance of a page for a crawler can also be expressed as a function of the similarity of a page to a given query. Different strategies are being employed in web crawling. These are as follows:-

A. Focused Web Crawler: Focused Crawler is the Web crawler that tries to download pages that are related to each other. It collects documents which are specific and relevant to the given topic.

B. Incremental Crawler: A traditional crawler, in order to refresh its collection, periodically replaces the old documents with the newly downloaded documents. On the contrary, an incremental crawler incrementally refreshes the existing collection of pages by visiting them frequently; based upon the estimate as to how often pages change.

C. Distributed Crawler: Distributed web crawling is a distributed computing technique. Many crawlers are working to distribute in the process of web crawling, in order to have the most coverage of the web. A central server manages the communication and synchronization of the nodes, as it is geographically distributed.

D. Parallel Crawler: Multiple crawlers are often run in parallel, which are referred as Parallel crawlers. A Parallel crawler can be on local network or be distributed at geographically distant locations .Parallelization of crawling system is very vital from the point of view of downloading documents in a reasonable amount of time.

III**. RELATED WORK**

In [1] used an effective harvesting framework for deep-web interface, namely Smart-Crawler. According to the used model this approach achieves both wide coverage for deep web interfaces and maintains highly efficient crawling.

In [2] major focus was on the fact that effective filters can be used to produce highly effective results on web. The filters incorporated with the used algorithms in the paper are well effective and high performance for web search, reduce the network traffic and crawling costs.

The work in [5] focused on using query preproccessing using fuzzy logic and also suggested that the query-based mechanism is based on the query scope, a measure of the query specificity. The query scope is dined using probabilistic propagation mechanism on top of the hierarchical structure of concepts provided by Word Net.

Also [6] focused that predictors can be generated before the retrieval process takes place, which is more practical than current approaches to query performance prediction. The approach was measured with the linear and non-parametric correlations of the predictors with Average Precision
The work in [7] used a model that focused on selective pruning framework for ensuring efficient yet effective retrieval, by appropriately setting the pruning parameters of Wand on a per-query basis, before re-ranking the results using a learned model.

IV. PROBLEM IDENTIFICATION

The amount of data consumed by crawler while searching is huge. The crawler searches large amount of data that may contain lots of irrelevant information. Also a lot of time is wasted for searching relevant data among the huge amount of irrelevant results returned by crawler and user has to waste a time while crawling on web while scanning irrelevant links also. Pre/Post query processing approaches and site-based searching approach can be integrated in order to pre-processing the user query. By integration of different processing approaches and link ranking approaches a lot of valuable user time is saved. Post query approach may also filter out all irrelevant information which is not necessary according to the query which is been fired, and gives the expected results. For pre-query approaches certain guidelines can be integrated with fuzzy logic:

- Implicit AND: User need not include the logical operator.
- Exact Matching: The length of the query.
- Word Variation: Provide user with option of search term variation.
- Common Word Exclusion: User need not include Stop Words like on , where and how.
- 32 Word Limit: Limit the user query length to 32 words.
- Case Sensitivity: Generally user queries are case insensitive.
- Ignoring Punctuation: User must not include characters like?,[]()@

V. PROPOSED WORK

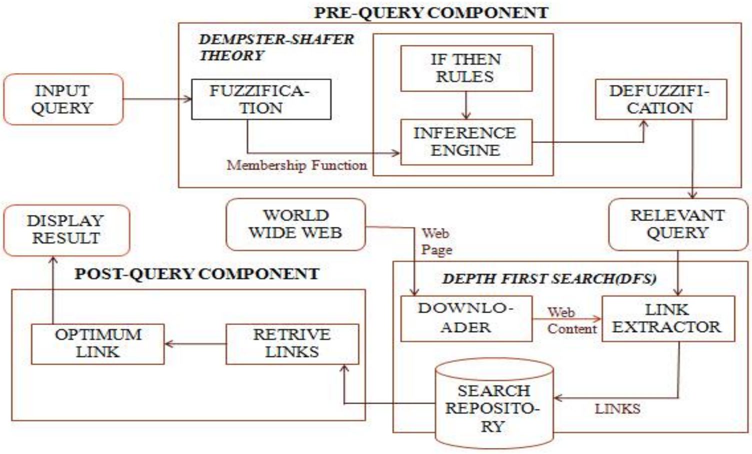

Figure 1: Efficient Crawling Technique

The basic approach consists of following steps:

- The query from the user is feeded to Pre-Query module. The relevance of the input query is first checked by inference engine in the pre-query component.
- Using the concept of fuzzy logic implemented by using IF-Then rules the result of the query relevance is returned to the user.
- The relevant query is further processed by using an appropriate Web-Crawling algorithm.
- The result of the crawling process is then provided to post-query module for filtering most relevant data.
- The final result is then displayed in front of the user.

The block diagram of proposes approach for smart crawling is given in Figure 1.

The basic algorithm is outlined as follows:-

- User will input query.
- Redirect query to the pre-query component.
- Apply Dempster–Shafer theory (DST).The theory allows one to combine evidence from different sources and arrive at a degree of belief that takes into account all the available evidence. The Dempster-Shafer Theory is a method of inexact reasoning.

Formula:

$$K = \Sigma\ m1\ (B).m2(C)$$

Where, K = measure conflict between two set of masses.
m1, m2 = two set masses
B, C = two different sets

Perform Fuzzification using IF-Then Rules

EX: Set MassA = containsAND(Query)
Set MassB = containsCommonWords (Query)
Set MassC = lengthGreaterThan32 (Query)
Set MassD = containsPunctuation(Query)
Set MassE = containsUpperAndLower(Query)

- If the query probability value is greater than or equal to 0.7 then it is relevant query and then it goes to next step or else if it irrelevant then directly while display a message to user for query irrelevancy.

Formula used:
Output Query is relevant if Probability >=0.7 else Irrelevant

- Through inference engine find the probability sets. And check the relevancy of the query.

Formulation used is:
Set K = MassA+MassB+MassC+MassD+MassE
Where, MassA ∩ MassB ∩ MassC ∩ MassD ∩ MassE = Φ
Set Probability = (1/ (1- K))*(MassA+MassB+MassC+MassD+MassE)

Where, K = measure of conflict between states
Mass N = different states.

- Perform crawling using DFS (Depth First Search) algorithm.
- Display result of query after Post-Query processing

The Web crawler has been currently implemented using DFS algorithm and results are obtained. The Depth first search algorithm is a more useful search which starts at the root URL and traverse depth through the child URL. First, it move to the left most child if one or more than one child exist and traverse deep until no more is available. Here backtracking is used to the next unvisited node and processes are repaid in similar manner. By the use of this algorithms it makes sure that all the edges, i.e. all URL is visited atleast once. It is very efficient for search problems. It can be used in cell phone so that it make searching process easy as it can directly filter out any give expected result. As this application can be easily work in many different languages, so that it become easy for the people who are unable to understand web searching processes .This approach can also be implemented for advertising.

REFERENCES

1. SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web, Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang & Hai Jin, Interfaces .IEEE Transactions on Services Computing Volume: PP Year: 2015.
2. Survey of Web Crawling Algorithms - Rahul kumar, Anurag Jain and Chetan Agrawal. Advances in Vision Computing: An International Journal (AVC) Vol.1, No.2/3, September 2014.
3. Dempster-Shafer theory for a query-biased combination of evidence on the Web- Vassilis Plachouras, Iadh Ounis. Springer-Verlag Berlin Heidelberg 2014.

4. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence, Ying Zhao, Falk Scholer, and Yohannes Tsegay, C.Macdonald et al. (Eds.): ECIR 2008, LNCS 4956, pp. 52–64, 2008._c Springer-Verlag Berlin Heidelberg 2008.
5. Inferring Query Performance Using Pre-retrieval Predictors, Ben He and Iadh Ounis, Department of Computing Science University of Glasgow fben,ounisg@dcs.gla.ac.uk.
6. A Unified Framework for Post-Retrieval Query-Performance Prediction, Oren Kurland, Anna Shtok, David Carmel, and Shay Hummel, ICTIR 2011, LNCS 6931, pp. 15–26, 2011. c_Springer-Verlag Berlin Heidelberg 2011.
7. Varying Approaches to Topical Web Query Classification , Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury,& Ophir Frieder, SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands, ACM ..
8. Survey on – Self Adaptive Focused Crawler, Ms. Pallavi Wadibhasme, & Prof. Nitin Shivale, Pallavi Wadibhasme et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 218-220 .
9. Efficient Query Evaluation using a Two-Level Retrieval Process- Andrei Z. Broder, David Carmel, Michael Herscovici, Aya Soffer & Jason Zien.
10. Evaluating Topic DrivenWeb Crawlers, Filippo Menczer, Gautam Pant,& Padmini Srinivasan