Deep Information Retrieval system by using N-Gram Approach

Pooja Ahuja¹, Pratima Sharma²
¹Assistant Professor, Computer Science Department, DAVIM, Faridabad, INDIA
²Assistant Professor, Computer Science Department, AITM, Palwal, INDIA

ABSTRACT

Information Retrieval is an activity to get information from an exhaustible collection of information resources. In this paper we are proposing a technique for improving the retrieval ratio of only relevant data from web. Earlier various techniques are already implemented using different approaches but our technique implies a n-gram approach on URL’s so that the we can provide the useful information as per user requirements while satisfying constraints.

Keywords: DIR, n-gram, query interface, crawlers.

I. INTRODUCTION

The concept of the deep Web is becoming more complex as search engines have found ways to integrate deep Web content into their central search function. This includes everything from airline flights to news to stock quotations to addresses to maps to activities on Facebook accounts. The Deep Web refers to content hidden behind HTML forms. In order to get to such content, a user has to perform a form submission with valid input values. The Deep Web has been acknowledged as a significant gap in the coverage of search engines because web crawlers employed by search engines rely on hyperlinks to discover new web pages and typically lack the ability to perform such form submissions. Various accounts have hypothesized that the Deep Web has an order of magnitude more data than the currently searchable World Wide Web. Furthermore, the Deep Web has been a long-standing challenge for the database community because it represents a large fraction of the structured data on the Web.

II. BACKGROUND

Balakrishnan [9], proposed the solution for accessing the deep web that solution named as “Source Rank”[9]. He worked to overcome critical challenge in searching the open collections like the deep web is identifying the important and the trustworthy sources, he evolved with a method named as SourceRank to evaluate the importance and the trustworthiness of the deep web sources based on the agreement of their solutions. He solved the detrimental problems of domain identification of sources, query routing, and self-adaption in both vertical and horizontal integration. The resultant solution, analyzes the results returned by sources, and interprets their degree of agreement as an implicit endorsement across sources. He performed horizontal and vertical integration of deep web by exploring the data both vertically and horizontally.

According to the Wei Liu, Xiao Feng Meng[10] gave a vision based approach for Deep web data extraction. World Wide Web has more and more online Web databases which can be searched through their Web query interfaces. The number of Web databases has reached 25 millions according to a recent survey. All the Web databases make up the Deep Web (hidden Web or invisible Web). Often the retrieved information (query results) is encapsulated in Web pages in the form of data records. These special Web pages are generated dynamically and are hard to index as compared to by traditional crawler based search engines, such as Google and Yahoo.

He made various contributions which are as follows:
1) A novel technique to perform data extraction from deep Web pages using primarily visual features with the utility to extract deep Web data automatically.
2) A new measure known as revision, is introduced to evaluate Web data extraction tools so that it can generate a perfect wrapper for a site.
3) A rigorous testing is performed on large data set consisting of 1,000 Web databases across 42 domains is used to give an efficient result.

Madhavan & Afanasiev[11] focussed on two major approaches for exploring the Deep web: the Virtual integration and Surfacing. He designed an algorithm for surfacing in order to minimize the surfaced page and maximizing coverage. They both implies that the extracted pages should neither have too many results on a single surfaced page nor too few. The Proposed algorithm selects a surfacing scheme which tries to ensure such an index ability criterion while minimizing the surfaced pages and maximizing coverage.

Ajoudanian, and Davarpanah Jazi[12] designed a system capable of extracting and matching information in the deep web. This system uses correlation approach to perform the matching attributes in online databases. This system accomplishes in two
essential steps enumerated as follows: a. Extracting information from query interfaces
b. Perform matching.

He and et. Al[12], conducted the survey for accessing the deep web and concluded that the deep Web is of a large scale of 307,000 sites, 450,000 databases, and 1,258,000 interface.

Raghavan and Molina[13] solved the problem of making a hidden Web crawler; with the functionality of crawling and extracting content from these hidden databases. It enables indexing, analysis, and mining of hidden Web content, akin to what is currently being achieved with the PIW with the ability to extract content by such crawlers in order to categorize and classify the hidden databases. They also built a prototype hidden Web crawler called HiWE (Hidden Web Exposer). And LITE (Layout Information Extraction Model)

III. PROPOSED METHODOLOGY

This paper proposing a technique in which the work is accomplished in 3 basic phases that are as follows:-
1. Creation of Query interface
2. Evaluating the URL ‘s relevancy by using N-gram approach
3. Continuous updation & Refreshal

IV. EVALUATION

This phase is the last but crucial phase as it directly deals with the information that provided to the user. The tasks of this phase is accomplished by the N-Gram approach. It works by matching the URL using BI-Gram. For ex: http://www.redwagnor.com

We will exclude http by setting delimiter (exclusion of http://www.)
Bi-Gram working illustrated below:

“If it matches then it generates the relevant information and image to the users, which is the main purpose. The working of the project is elaborated by the following in figure 2

V. CONCLUSION

The work concludes that retrieval is done in basic two phases. First is creating a Query interface and retrieving the Relevant URL’s then applying N-gram approach on URL’s and then opening the relevant URL to show the information satisfying users constraints as well as attributes.

VI. FUTURE SCOPE

One critical challenge is source and query classification to different domains, and routing the queries to the best sources as well as retaining and utilizing the semantics of exposed content are interesting problems for future research. Second challenge of estimating coverage is also relevant in the context of database exploration and remains an open problem. Third challenge tries to ensure such an index ability criterion while also minimizing the surfaced pages and maximizing coverage. Fourth challenge is to adapt the search to the dynamism of the deep web, like the changes in the interfaces, data quality, record structure, response time etc, since these ubiquitous changes often break down the search.

REFERENCES


[10] Wei Liu, Xiaofeng Meng, Member, IEEE, and Weiyi Meng, Member, IEEE. VIDE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. X, XXXXXXX 2010.

