The rapid development of computer and networking technologies has increased the popularity of the web, which has led to the presence of more and more information on the web. However, the explosive increase of information online leads to some search problems—specifically search engines usually return too many unrelated results on a given query.

Deep web is content that is dynamically generated from data sources, namely file systems or databases. Unlike the surface web, pages in the deep web are collected by following the hyperlinks embedded within collected pages. Data from the deep web are guarded by search interfaces such as web services, HTML forms or programmable web, and they can be retrieved by database queries only. Surface web content is defined as static, crawlable web pages. The surface web contains a large amount of unfiltered information, whereas the deep web includes high-quality, managed and subject-specific information. The deep web grows faster than the surface web because the surface web is limited to what is easily found by search engines.

The deep web covers domains such as education, sports and the economy. It contains huge amounts of information and valuable content. Because deep web information can be found only by queries, it is necessary to design a special search engine to crawl deep web pages. Deep web data extraction is the process of extracting a set of data records and the items that they contain from a query result page. Such structured data can be later integrated into results from other data sources and given to the user in a single, cohesive view. Domain identification is used to identify the query interfaces related to the domain from the forms obtained in the search process. The domain classifications are done based on the number of matching results obtained in the similar criteria among the query interface and the domain, based on the database summary.

**DWDE FRAMEWORK BASED ON URL AND DOMAIN CLASSIFICATION**

The Deep Web Data Extraction (DWDE) framework seeks to provide accurate results to users based on their URL or domain search. The complete steps of the framework for DWDE are shown in figure 1. Initially, the collected web sites are categorized into surface web or deep web repositories based on their content. The user gives a query to retrieve the relevant web pages. The user can search the query based on the following two criteria:

1. URL
2. Domain

If the user searches by URL, then the proposed framework validates whether it is a live web site. If it is a valid web site, the necessary and important contents are extracted from the web site based on the tag information. If the keyword or domain information is directly given to the search engine, the contents are extracted based on the given keyword and web site content matching.

For both searching criteria, a frequency calculation is applied to calculate the number of occurrences among the given query with the relevant web sites. The domain classification algorithm is designed to predict the classified domains, and it retrieves the accurate web pages for users.

**Classifying the Web Site**

The Internet contains a huge number of web pages and web content. Web pages can be categorized into two types, namely surface web repository and deep web repository. This classification is based on the static and dynamic nature of the web pages. The web
sites are classified based on the tag information. If the web site includes any form tags, it is classified as a deep web site. (It includes dynamic information.) If the web site does not include any form tags, it is classified as surface web repositories. (It includes only static information.) The DWDE framework helps process deep web pages to provide the accurate and necessary web pages to the users.

**Tag-based Feature Extraction (TFE)**
Tags, such as title, header, anchor, metadata about the Hypertext Markup Language (HTML) document, paragraph, group in-line elements in a document and images, are utilized to extract the feature’s domain classification. Most of the domain-specific necessary terms appear under these tags. The existing web-page classification mechanism uses the HTML tags for domain classification. In this classification method, a stop word removal is applied to all the information from each of those tags to extract the essential features. Each stemmed term with corresponding tag creates a feature. For example, the word “mining” in the title tag, the word “mining” in the <b> tag and the word “mining” in the <li> tag are all considered different features from similar HTML tags.

The DWDE framework uses limited tags to extract the most important features to recognize the domain of the given URL. These tags are used to avoid spending time extracting the less-important features.

**Domain Classification**
After extracting the source information, a visual block tree is constructed using the tags. Constructing a tag tree from the web site is an essential step for most web content extraction methods. In the DWDE framework, the HTML tag is used to formulate the corresponding tag tree. Based on the properties of the HTML tag and text, the tag node is defined as tag-name, type, parent, child-list, data, text-num and attribute. The tag-name denotes the name of the tag; type denotes the type of each node and where nodes are divided into branch nodes; parent represents the parent node; child-list is the set of successors; data stores the content of the node; text-num denotes the total number of punctuation and words in all the descendants of each node; and attribute denotes a mapping of the characteristics of the HTML tag. The root node represents the whole page, and each block in the tree belongs to the clocks that cannot be further segmented. For each tag on the tree, keyword frequency is calculated for the extracted information.

**PERFORMANCE ANALYSIS**
The DWDE framework was compared with the existing Genetic Algorithm (GA)\(^3\) and Naive Bayes (NB)\(^4\) classifiers with respect to precision, recall and F-measure. The proposed system is able to search the query with multiple keywords.
so it is also compared with the existing Multi-keyword Text Search (MTS)\(^5\) algorithm, and the execution time is investigated. The DWDE framework can be executed for any database that collects from real-time online data sets.

**Precision and Recall Analysis**

Precision is the number of true positives divided by the total number of positives, providing the percentage of true positives. Recall is the number of true positives divided by the number of true positives and false negatives, providing the percentage of positives that are found.

The greater the value of the F-measure, the better the performance of the system. The DWDE framework results in higher F-measure values than the GA and NB methods. As the precision and recall values for the method yield better results, it is reflected in the F-measure calculation. The DWDE framework can result in better classification of web pages than the existing methods. The result of the F-measure analysis among various domains is depicted in **figure 4**.

In this experiment, the DWDE framework was validated for diverse domains, including association, radio, travel, disaster and child games. The precision and recall values are noted and the results are shown in **figures 2** and **3**. The proposed DWDE framework uses only a limited number of valuable tags to classify the domains, so the results are better and more relevant pages for the user. As a result, the framework automatically has better precision and recall values.

The results show that the DWDE framework results in better performance than the existing GA and NB methods.

**F-measure**

F-measure is the measure of a system’s/method’s performance by its classifiers and is based on precision and recall scores.
Execution Time Analysis

The execution time is evaluated based upon three types of processes:
- Time taken for the raw data set
- Time taken for HTML parsing
- Time taken for domain classification

The three criteria are investigated for the proposed method with the existing classifiers. The DWDE framework uses limited tags to extract the features, so it takes less time to execute the domain classification process to retrieve the resultant web pages.

The results (figure 5) show that the proposed framework takes less execution time to complete the task than the GA and NB methods.

Multi-keyword search is also possible in the DWDE framework. The execution time experiment was conducted for multiple keywords (up to five keywords). The DWDE method takes less execution time than the existing MTS algorithm, which is shown in figure 6.

CONCLUSION

The DWDE framework uses the tag-based feature extraction algorithm to retrieve the necessary data. The method can process the query in two ways, searching by URL and searching by domain. Hence, it provides a user-friendly search process to deliver the results. Most of the existing algorithms process the queries on single query or multiple keyword queries, but the DWDE framework can process the single, multiple keyword queries and appropriate domain classification.

The frequency calculation is applied to compare the matching frequencies between the user query and the relevant search web sites. Based on the frequency measures, the domain classification algorithm is introduced to retrieve the accurate resultant pages.

The experiment results are compared with the existing methods, such as GA, NB and MTS algorithm. The proposed framework has better precision, recall and F-measure values than the existing GA and NB classifiers. Moreover, the time taken to execute the query search is less than the existing MTS algorithm. The DWDE framework is well suited to search the query for efficient user query retrieval.

ENDNOTES

4 Ibid.

5 Sun, W.; B. Wang; N. Cao; M. Li; W. Lou; Y. Hou; “Verifiable Privacy-preserving Multi-keyword Text Search in the Cloud Supporting Similarity-based Ranking,” IEEE Transactions on Parallel and Distributed Systems, 2015