Chapter 2
Web Data Extraction and Integration System for Search Engine Results

K. R. Venugopal and K. C. Srikantaiah

Abstract There is an explosive growth of information in the World Wide Web thus posing a challenge to Web users to extract essential knowledge from the Web. Search engines help us to narrow down the search in the form of Search Engine Result Pages (SERP). Web Content Mining is one of the techniques that help users to extract useful information from these SERPs. In this chapter, we propose two similarity-based mechanisms; Web Data Extraction using Similarity Function (WDES), to extract desired SERPs and store them in the local depository for offline browsing and Web Data Integration using Cosine Similarity (WDICS), to integrate the requested contents and enable the user to perform the intended analysis and extract the desired information. Our experimental results show that WDES and WDICS outperform Data Extraction based Partial Tree Alignment (DEPTA) [1] in terms of Precision and Recall.

2.1 Introduction

The World Wide Web (WWW) has now become the largest knowledge base in human history. The Web encourages decentralized authorizing in which users can create or modify documents locally, which makes information publishing more convenient and faster than ever. Because of these characteristics, the Internet has grown rapidly, which creates a new and huge media for information sharing and exchange. Most of the information on the Internet cannot be directly accessed *via* the static link,

©Reprinted by permission from Springer Nature: Springer, Berlin, Heidelberg, LNICST-108, K. C. Srikantaiah, M. Suraj, K. R. Venugopal, S. S. Iyengar, L. M. Patnaik, Similarity based Web data extraction and integration system for Web content mining, pages: 269–274, 2012.

K. R. Venugopal (✉)
Bangalore University, Jnana Bharathi, Bengaluru 560056, India
e-mail: venugopalkr@gmail.com

K. C. Srikantaiah
SJB Institute of Technology, BGS Health and Education City, Uttarahalli Main Road, Kengeri, Bengaluru 560060, India
e-mail: srikantaiahkc@gmail.com

K. R. Venugopal et al., *Web Recommendations Systems*,
https://doi.org/10.1007/978-981-15-2513-1_2

must use Keywords and Search Engine. Web search engines are programs used to search information on the WWW and FTP servers that check the accuracy of the data automatically. When searching for a topic in the WWW, it returns many links or Web sites related, i.e. Search Engine Result Pages (SERP) on the browser to a given topic. Some data on the Internet that is visible to the search engine is called surface Web, whereas some data such as dynamic data in dynamic database invisible to the search engine is called deep Web.

There are situations in which the user needs those Web pages on the Internet to be available offline for convenience. The reason being offline availability of data, limited download slots, storing data for future use, etc. This essentially leads to downloading raw data from the Web pages on the Internet, which is a major set of the inputs to a variety of software that are available today for the purpose of data mining. Web data extraction is the process of extracting the information that users are interested in, from semi-structured or unstructured Web pages and saving the information as the XML document. During Web data extraction phase, search engine result pages are crawled and stored in the local repository. Web database integration is a process of extracting the required data from the Web pages stored in the local repository and integration of the extracted data that needs to be stored in the database.

In recent years, there have been rapid improvements on technology with products differing in the slightest of terms. Every product needs to be tested thoroughly and the Internet plays a vital role in gathering information for the effective analysis of the products. In our approach, we replicate search engine result pages locally based on comparing page URLs with a predefined threshold. The replication is such that the pages are accessible locally in the same manner as on the Web. In order to make the data available locally to the user for analysis, we extract and integrate the data based on the prerequisites which are defined in the configuration file.

In a given set of Web pages, it is difficult to extract matching data. So, we have to develop a tool that is capable of extracting the exact data from the Web pages. In this chapter, we have developed the WDES algorithm, which provides offline browsing of the pages. Here, we integrate the downloaded content onto a defined database and provide a platform for efficient mining of the data required.

2.2 Related Works

Zhai et al. [1] propose Data Extraction based Partial Tree Alignment (DEPTA) algorithm for the structured data extraction from the Web based on partial tree alignment, and structured data extraction from arbitrary Web pages. The main objective is to automatically segment data records in a page, extract data items/fields from these records and store the extracted data in a database. It consists of two steps, i.e. (i) identifying individual records in a page using visual information and (ii) aligning and extracting data items from the identified records using tree matching and a novel partial alignment technique, respectively.

Ananthanarayanan et al. [2] propose a method for offline Web browsing that is minimally dependent on real-time network availability. The approach makes use of the Really Simple Syndication (RSS) feeds from Web servers and prefetches all new content specified, defining the content section of the home page. It features intelligent prefetching, robust and resilient measures for intermittent network handling, template identifier and local stitching of the dynamic content into the template. It does not provide all the information on the page. The drawback is, the content defined in the RSS feeds may not be updated nor they do provide for dynamic changes in the page.

Myllymaki et al. [3] describe ANDES, a software framework that provides a platform for building a production-quality Web data extraction process. The key aspects are that it uses XML technology for data extraction, including XHTML and XSLT and provides access to the deep Web. It addresses the issues of website navigation, data extraction, structure synthesis, data mapping and data integration. The framework shows that production-quality Web data extraction is quite feasible and that incorporating domain knowledge into the data extraction process can be effective in ensuring the high quality of extracted data. Data validation technique, a cross-system support, and a well-established framework has not been addressed.

Yang et al. [4] propose a novel model to extract data from deep Web pages. It consists of four layers, among which, the access schedule, extraction layer and data cleaner are based on the rules of structure, logic and application. The model first uses the two regularities of the domain knowledge and interface similarity to assign the tasks that are proposed from the users and chooses the most effective set of sites to visit. Then, the model extracts and corrects the data based on the logical rules and structural characteristics of the acquired pages. Finally, it cleans and orders the raw dataset to adapt to the customs of the application layer for later use by its interface.

Yin et al. [5] propose Web page templates and DOM technology to effectively extract simple structured information from the Web. The main contents include the methods based on edit distance, DOM document similarity judgement, clustering methods of Web page templates and programming an information extraction engine. The method provides steps for information extraction using DOM tree parsing and on how a page similarity judgement is to be made. The template extraction and reconstruction is depicted in order to find out how the data has been parsed on the Web page and in reconstructing the page to overcome the noise on the page. It does not parse through the dynamic content of scripts on the page.

Liu et al. [6] propose a method to extract the main text from the body of a page based on the position of DIV. It reconstructs and analyses DIV in a Web page by completely simulating its display in a browser, without additional information such as Web page template; its implementation complexity is quite low. The core idea includes the concept of atomic DIV, i.e. a DIV block that does not include other DIVs. Then it filters out redundancy and reconstructs by clustering, reposition analysis and finally stores the elements of the data in an array. The selected DIVs in the selected array contains the main text of the page. We can get the main text by combining these DIVs. The method has high versatility and accuracy. The majority of the Web page content has been made up of tables and this approach does not address the table layout data. This drastically reduces the accuracy of the entire system.

Dalvi et al. [7] explore a novel approach based on temporal snapshots of Web pages to develop a tree-edit model of HTML to improve wrapper construction. The model is attractive as a source tree transformed into a target tree can be estimated efficiently, in quadratic time, making it a potentially useful tool for a variety of tree-evolution problems. An improvement in the robustness and performance and ways to prune the trees without hampering model quality is to be dealt with.

Novotny et al. [8] represent a chain of techniques for the extraction of object attribute data from the Web pages. They discover data regions containing multiple data records and also provide a detailed algorithm for detail page extraction based on the comparison of two html subtrees. They describe the extraction from two kinds of Web pages: master pages containing structured data about multiple objects and detail pages containing data about single product, respectively. They combine the techniques of the master page extraction algorithm, detail page extraction algorithm and comparison of sources of two Web pages. The approach makes use of attribute values based on the Document Object Model (DOM) structure described in the Web pages. It has better precision of extraction of values from pages defined in the master and detailed format and also minimizes the user effort. Enhancements to the approach may include the page-level complexity of having multiple interleaved detail pages to be traversed, coagulation of different segments in the master page and series implementation of pages.

Nie et al. [9] provide an approach for obtaining data from the result set pages by crawling through the pages for data extraction based on the classification of the URL (Unified Resource Locator). It extracts data from the pages by computing the similarity between the URLs of hyperlinks and classifying them into four categories, where each category maps into a set of similar Web pages, which separates result pages from others. It makes use of the page probing method to verify the correctness of classification and improves the accuracy of crawled pages. The approach makes use of the minimum edit distance algorithm and URL-field algorithm to calculate the similarity between URLs of hyperlinks, respectively. However, there are a few constraints to this approach. It is not able to resolve issues of pages related to the partial page refreshments by the use of JavaScript engines.

Papadakis et al. [10] describe STAVIES, a novel system for information extraction from Web data source through automatic wrapper generation using clustering technique. The approach is to exploit the format of the Web pages to discover the underlying structure in order to finally infer and extract pieces of information from the Web page. The system can operate without any human intervention and does not require any training.

Chang et al. [11] have studied the major Web data extraction systems and compare them in three dimensions: the task domain, the automation degree and the techniques used. These approaches emphasize on availability of robust, flexible Information Extraction (IE) systems that transform the Web pages into program-friendly structures such as a relational database. It mainly focuses on the IE from semi-structured documents and discusses only those that have been used for Web data. The trend of highly automatic IE systems saves the effort of programming, labelling, enhancements for applying the techniques to non-html documents such as medical records and curriculum vitae and facilitates the maintenance of larger semi-structured documents.

Angkawattanawit et al. [12] propose an algorithm to improve harvest rate by utilizing several databases like seed URLs, topic keywords and URL relevance predictors that are built from previous crawl logs. Seed URLs are computed using BHITS [13] algorithm on previously found pages by selecting pages with high hub and authority scores that are used for future recrawls. The interested keywords for the target topic are extracted from the anchor tags and title of previously found relevant pages. Link crawl priority is computed as a weighted combination of popularity of the source page, similarity of link anchor text to topic keywords and predicted link score are based on previously seen relevance for that specific URL.

Aggarwal et al. [14] propose the concept of intelligent crawling where the user can specify an arbitrary predicate such as keywords, document similarity, etc., which are used to determine documents relevance to the crawl; the system adapts itself in order to maximize the harvest rate. A probabilistic model for URL priority prediction is trained using URL tokens, information about content of in-linking pages, number of sibling pages matching the predicate and short-range locality information

Chakrabarti et al. [15] propose models for finding URL visit priorities and page relevance. The model for URL ranking called apprentice is trained online by samples consisting of source page features and the relevance of the target page but the model for evaluating page relevance can be anything that outputs a binary classification. For each retrieved page, the apprentice is trained based on information from the baseline classifier and features around the link extracted from the parent page to predict the relevance of the page pointed to by the link. Those predictions are then used to order URLs in the crawl priority queue. The number of false positives has decreased significantly.

Ehrig et al. [16] propose an ontology-based algorithm for page relevance computation which is used for Web data extraction. After preprocessing, words occurring in the ontology are extracted from the page and counted. The relevance of the page with regards to user-selected entities of interest is then computed by using several measures such as direct match, taxonomic and more complex relationships on ontology graph. The harvest rate of this approach is better than baseline-focused crawler.

Srikantaiah et al. [17] propose an algorithm for Web data extraction and integration based on URL similarity and cosine Similarity. The extraction algorithm is used to crawl the relevant pages and store them in a local repository. The integration algorithm is used to integrate similar data in various records based on cosine similarity.

2.3 System Architecture

2.3.1 Problem Definition

Given a start page URL and a configuration file, the main objective is to extract pages which are hyperlinked from the start page and integrate the required data for analysis using data mining techniques. The user has sufficient space on the machine to store the data that is downloaded. The basic notations used in the model are shown in Table 2.1.

The entire system has been divided into three blocks that are able to accomplish the dedicated tasks, working independently from one another. The first block, namely the Web Data Extractor connects to the Internet to extract the Web pages described by the user and stores it onto a local repository. It also stores the data in a format that is likely to be an offline browsing system. Offline Browsing means that the user is able to browse through the pages that have been downloaded, by the use of a Web browser without having to connect to the Internet. Thus, it would be convenient for the user to go through the pages as and when he needs it.

The second block, namely the Database Integrator extracts the vital piece of information that the user needs from the downloaded Web pages and creates a database with the set of tables and respective attributes in the database. These tables are populated with the data extracted from the locally downloaded Web pages. This makes it easy for the user to query out his needs from the database.

The overall view of the system is as shown in Fig. 2.1. The system initializes a set of inputs for each block. These blocks are highlighted and framed according to the flow of data. The inputs that are needed for the start of the first block, i.e. the Web Data Extractor are fed in through the initialize phase. On receiving the inputs, the search engine performs its task of navigating to the page on the given URL and entering the criteria defined on the configuration file. This populates a page from which the actual download can start.

The process of extraction is to achieve the task of downloading the contents from the Web and to store them onto the local repository. Then onwards the process

Table 2.1 Basic notations

S	: Start Page URL	C, C_i	: Configuration File
l	: Depth of Recursion	W	: Set of Search Engine result Pages
$H(W)$	: Hyperlinks Set of W	CL	: Current Level
Lp	: Local Path to Hyperlinks	T_0	: Threshold for Similarity.

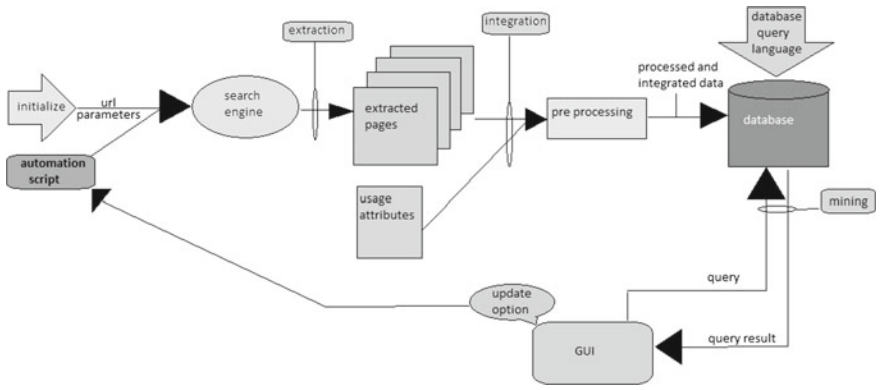


Fig. 2.1 Bluetooth SIG site

of extracting the pages iterates over, resulting in the outcome of offline Web pages (termed extracted pages). The extracted pages are available offline on a pre-described local repository. The pages in the repository can be browsed through in the same manner as that available on the Internet with the help of a Web browser. The noticeable thing here is that the pages are available offline, thereby are much faster to be accessed.

The outcome of the first block is to be chained with other attributes that are essential for the second block, namely the Database Integrator. The task of the second block is to extract the vital information from the extracted pages, process it and store it in accordance onto the database. The database is predefined in the set of attributes, together with the tables. This block needs the presence of the usage attributes that are extracted from the downloaded Web content. The usage attributes are mainly defined in a file termed as configuration file that the user needs to prepare before the execution. The file also contains the table to which the data extracted are to be added together with the table attributes and mapping.

On successful entries of the input, the integration block is able to accomplish the task of extracting the content from the Web pages. Since the Web pages do tend to remain in the same format as on the Internet, it is easy to be able to navigate across these pages as the links refer to the locally extracted pages. The extracted content is then processed to meet the desired data attributes that are listed on the database and the values are dumped onto it. This essentially creates rows of extracted information in the database. The outcome of second block results in tuples in the database that are essentially extracted out from the extracted pages got from the first block.

Now, that the database has been formed with the vital information contained in it, it would well be the task of the analyser, i.e. the third block referred to as GUI to provide the user with the functionality on how to deal with the contained data. The GUI provides for the interfaces that the user is able to achieve so as to obtain the data contained in the database, as and how he needs it. It acts as a miner that provides the user with the information obtained from the result of queries that are defined. The GUI also has options on referring the other blocks as requested by the user. This indeed interfaces all the three blocks, thereby providing the user with a better understanding and handling feature.

It is quite essential to know a few things that relate to the working of the entire system. First of all, it is very much essential that the inputs, initializations and the prerequisite data such as the usage attributes and/or configuration files that have been defined, do fall in a particular order and stick to the conventions defined on them. It is important to have the blocks to perform the tasks in order. The next thing is that, although the blocks do persist to function independently it is important that without the essential requirements, it is of no use to extract its functionality. Unless the prerequisites of the blocks are not met, the functionality that they accomplish is of no use. Similarly, to query out the needs of the user through the help of the GUI, it is essential that the data is available in the database. It is also essential that the first block and second block be run often and in unison so as to have an update on the current and ongoing dataset.

2.4 Mathematical Model and Algorithms

2.4.1 Web Data Extraction Using Similarity Function (WDES)

A connection has been established to the given *url* S and the page is processed with the parameters obtained from the configuration file C. On completion of this, we obtain the Web document that contains the links to all the desired contents that are obtained out of the search performed. The Web document contains individual sets of links that are displayed on each of the search results pages that are obtained. For example, if a search result obtained contains 150 records displayed as 10 records per page (in total 15 pages of information), we would have 15 sets of Web documents each containing 10 hyperlinks pointing to the required data. This forms the set of Web documents, W. That is,

$$W = \{w_i : 1 \le i \le n\}. \tag{2.1}$$

Each Web document $w_i \in W$ is read through to collect the hyperlinks that are contained in it, that are to be fetched to obtain the data values. We represent this hyperlink set as $H(W)$. Thus, we consider $H(W)$ as a whole set containing all the sets of hyperlinks on each page $w_i \in W$. That is,

$$H(W) = \{H(w_i) : 1 \le i \le n\}. \tag{2.2}$$

Then, considering each hyperlink $h_j \in H(w_i)$, we find the similarity between h_j and S, using Eq. 2.3.

$$SIM(h_j, S) = \frac{\sum_{i=1}^{min(nf(h_j),nf(S))} fsim(f_i h_j, f_i S)}{(nf(h_j) + nf(S))/2}. \tag{2.3}$$

where $nf(X)$ is the number of fields in X and $fsim(f_i h_j, f_i S)$ is defined as

$$fsim(f_i h_j, f_i S) = \begin{cases} 1\, if\ f_i h_j = f_i S \\ 0\, if\ f_i h_j \neq f_i S \end{cases} \tag{2.4}$$

The similarity $SIM(h_j,S)$ is the value that lies between 0 and 1, this value is used to compare with the defined threshold $T(0.25)$, we download the page corresponding to h_j to local repository if $SIM(h_j, S) \geq T$. The detailed algorithm of WDES is given in Algorithm 2.1.

Algorithm WDES navigates the search result page from the given $URL\ S$ and configuration file C and generates a set of Web documents W. Next, call the function $Hypcollection$ to collect hyperlinks of all pages in w_i, indexed by $H(w_i)$, page corresponding to $H(w_i)$ is stored in the local repository. The function $webextract$ is recursively called for each $H(w_i)$. Then, for each $h_i \in H(w_i)$, similarity between h_i and S is calculated using Eq. 2.3, if SIM(h_i, S) is greater than the threshold T_o, then page corresponding to h_i is stored and collect all the hyperlinks in h_i to X. Continue this process for X, until it reaches maximum depth l.

2.4.2 Web Data Integration Using Cosine Similarity (WDICS)

The objective of this algorithm is to extract data from the downloaded Web pages (those Web pages that are available in the local repository, i.e. output of WDES algorithm) into the database based on attributes and keywords from the configuration file C_i. We collect all the result pages W from local repository indexed by S, then $H(W)$ is obtained by collecting all hyperlinks from W, considering each hyperlink $h_j \in H(w_i)$ such that $k \in keywords$ in C_i. On existence of k in h_j, we populate the new record set $N[m, n]$ by passing page h_j and obtaining values defined with respect to the $attributes[n]$ in C_i. We then populate the old record set $O[m, n]$ by obtaining all values with respect to $attributes[n]$ in database.

For each record i, $1 \leq i \leq m$, we find the similarity between $N[i]$ and $O[i]$ using cosine similarity

$$SimRecord(N_i, O_i) = \frac{\sum_{j=1}^{n} N_{ij} O_{ij}}{\sqrt{\sum_{j=1}^{n} N_{ij}^2 \sum_{j=1}^{n} O_{ij}^2}}. \tag{2.5}$$

If similarity between records is equal to zero, then we compare each $attribute[j]$ $1 \leq j \leq n$ in the records and form $IntegratedData$ with use of $Union$ operation and store in the database.

$$IntegratedData = Union(N_{ij}, O_{ij}). \tag{2.6}$$

The detailed algorithm of WDICS is shown in Algorithm 2.2. The Algorithms WDES and WDICS, respectively, extract and integrate data in Depth First Search (DFS) manner. Hence, their complexity is O(n^2), where n is the number of hyperlinks in H(W).

Algorithm 2.1: Web Data Extraction using Similarity Function (WDES)

```
Input  : $S$: Starting Page URL, $C$: Parameter Configuration File, $l$: Level of Data
         Extraction, $T_o$: Threshold.
Output : Set of Webpages in Local Repository.
begin
    $W$=Navigate to Web document on Given $S$ and automate page with $C$
    $H(W)$=Call: Hypcollection(W)
    for each $H(w_i) \in H(w)$ do
        Save page $H(w_i)$ on local Machine page $P$
        Call: Webextract($H(w_i)$, 0, pageppath)
Function Hypcollection($W$)
begin
    for each $w_i \in W$ do
        $H(w_i)$=Collect all hyperlinks in $w_i$
    return $H(W)$
Function Webextract($Z$, $cl$, $lp$)
Input  : $Z$: set of URLs, $cl$: Current level, $lp$: local path to Z.
Output : Set of Webpages in Local Repository.
begin
    for each $h_i \in Z$ do
        if $SIM(h_i, S) > T_o$ then
            Save $h_i$ to $F_{h_i}$
            $X$=collect URLs from $h_i$ and change its path in $lp$
            if ($cl < l$) then
                Call: Webextract($X$, $cl + 1$, pageppath of $X$)
```

2.5 Experiments

The algorithms are independent of usage and that they could be used on any given platform and dataset. The experiment was conducted on the Bluetooth SIG Website [18], The Bluetooth SIG (Special Interest Group) is a body that oversees the licensing and qualification of Bluetooth-enabled product. It maintains the database of all the qualified products in a database which can be accessed through queries on its website and hence is a domain-specific search engine. The qualification data contains a lot of parameters including company name, product name, any previous qualification

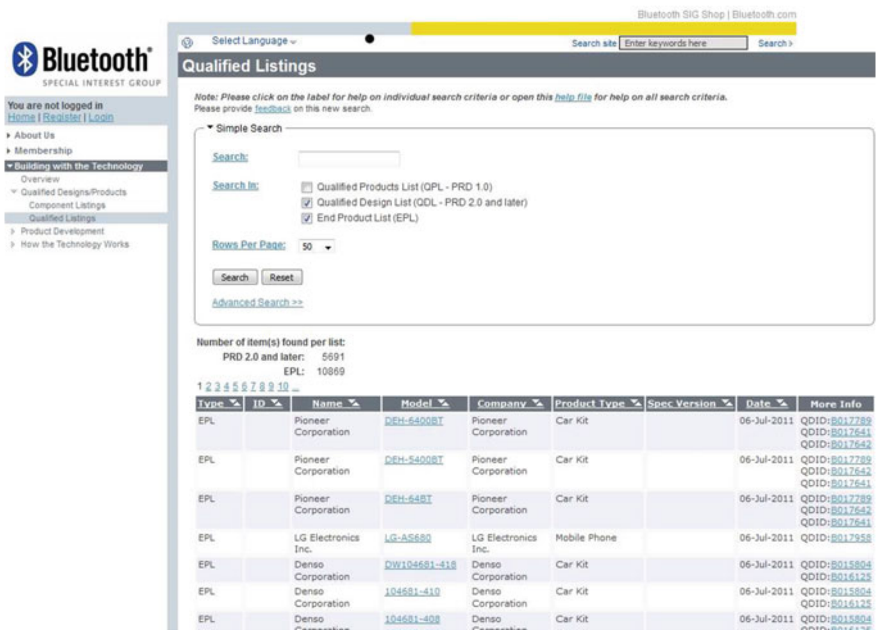
Algorithm 2.2: Web Data Integration using Cosine Similarity (WDICS)

```
Input  : $S$: Starting Page URL stored in local repository (output of WDES),
         $C_i$: Configuration File (Attributes and Keywords).
Output : Integrated Data in Local Repository.
1  begin
2      $H(w)$=Call: Hypcollection($S$)
3      for each $H(w_i) \in H(w)$ do do
4          Call: Integrate($H(w_i)$)
5  Function Integrate(X)
   Input  : $X$: set of URLs
   Output : Integration of Values of Attributes Local Repository
6  begin
7      for each $h_i \in Z$ do
8          if ($h_i$ contain keyword) then
9              $new[m][n]$=parse page to obtain values of defined attributes[n] in $C_i$
10             $old[m][n]$=obtain all values of attribute[n] from repository
11             for each record $i$ do do
12                 if ($SimRecord(new[i], old[i])$==1) then
13                     Skip
14                 else
15                     for each attribute $j$ do
16                         if ( $new[i][j]$ Not Equal to $old[i][j]$ ) then
17                             $Integrateddata$=union($new[i][j]$,$old[i][j]$)
18                     store $Integarteddata$ in local repository
19                 $X$=collect all links for $h_i$
20                 if ($X$ not equal to NULL) then
21                     Call: Integrate($X$)
```

used, etc. The main tasks for the Bluetooth SIG are to publish Bluetooth specifications, administer the qualification program, protect the Bluetooth trademarks and evangelize Bluetooth wireless technology. The key idea behind the approach is to collect and automate the collection of competitive and market information from the Bluetooth SIG site.

The site contains data in the form of a list that is displayed on the startup page. The display is formed based on the three types of listings; PRD 1.0, PRD 2.0 and EPL. The PRD 1.0 and PRD 2.0 are the qualified products and design list and EPL are the end product list being displayed. Each of the PRDs listed here are products that contain the specifications involved in them. The EPL are those that are formed in unison of the PRDs. Each PRD is identified by a QDID (an id for uniqueness) and each EPL is identified by the model. Each PRD may have many numbers of related EPLs and each EPL may have many numbers of related PRDs.

The listings as shown in Fig. 2.2, which contain links which navigate to the detailed content of the products that have been displayed here. The navigations on the page



Qualified Listings

Note: Please click on the label for help on individual search criteria or open this help file for help on all search criteria.
Please provide feedback on this new search.

Simple Search

Search:

Search In: [ ] Qualified Products List (QPL - PRD 1.0) [✓ Qualified Design List (QDL - PRD 2.0 and later)] [✓ End Product List (EPL)]

Rows Per Page: 50

Search Reset

Advanced Search >>

Number of item(s) found per list:
PRD 2.0 and later: 5691
EPL: 10869

1 2 3 4 5 6 7 8 9 10 ...

Type	ID	Name	Model	Company	Product Type	Spec Version	Date	More Info
EPL		Pioneer Corporation	DEH-6400BT	Pioneer Corporation	Car Kit		06-Jul-2011	QDID:B017789 QDID:B017641 QDID:B017642
EPL		Pioneer Corporation	DEH-5400BT	Pioneer Corporation	Car Kit		06-Jul-2011	QDID:B017789 QDID:B017642 QDID:B017641
EPL		Pioneer Corporation	DEH-64BT	Pioneer Corporation	Car Kit		06-Jul-2011	QDID:B017789 QDID:B017642 QDID:B017641
EPL		LG Electronics Inc.	LG-A5680	LG Electronics Inc.	Mobile Phone		06-Jul-2011	QDID:B017958
EPL		Denso Corporation	DW104681-418	Denso Corporation	Car Kit		06-Jul-2011	QDID:B015804 QDID:B016125
EPL		Denso Corporation	104681-410	Denso Corporation	Car Kit		06-Jul-2011	QDID:B015804 QDID:B016125
EPL		Denso	104681-408	Denso	Car Kit		06-Jul-2011	QDID:B015804

Fig. 2.2 SERP in bluetooth SIG site

reach to N number of pages before the case of termination. Thereby we may want to parse across the links to reach all the places to obtain the required data. Here, although we have all the data being displayed on the website, it is still not possible for us to prepare an analysis report based on the same. We want to play around with the data to get it to the form that we deserve it to be. Thereby, we incorporate the use of our algorithms to extract, integrate and mine the data from this site.

The experimental setup involves a Pentium Core 2 Duo machine with 2 GB RAM running windows. The algorithms have been implemented using a Java JDK and JRE 1.6, Java Parser with an access to an SQLite database and active broadband connection. Data has been collected from www.bluetooth.org, which lists the qualified products of Bluetooth devices. We have extracted 92 pages, with each page containing 200 records of information and data extraction was possible from each of these pages. Hence, we have a cumulative set of data for comparison based on the data extracted on the given attribute mentioned in the configuration file.

2.5.1 Precision and Recall Versus Attributes

Precision is defined as the ratio of correct pages and extracted pages and recall is defined as the ratio of extracted pages and the total number of pages. They are

Table 2.2 Performance evaluation between WDICS and DEPTA

Attributes	Total records (TR)	DEPTA		WDICS	
		Extracted records (ER)	Correct records (CR)	Extracted records (ER)	Correct records (CR)
Name	18234	18204	17325	18234	18234
Model	18234	17860	17010	18060	18060
Company	18234	18095	17208	18234	18198
Spec version	5508	5410	5016	5508	5426
Product type	18234	17834	17015	18234	18045
Total	78444	77403	73574	78270	77963
Recall=(ER/TR)*100		98.67%		99.77%	
Precision=(CR/ER)*100		95.05%		99.60%	

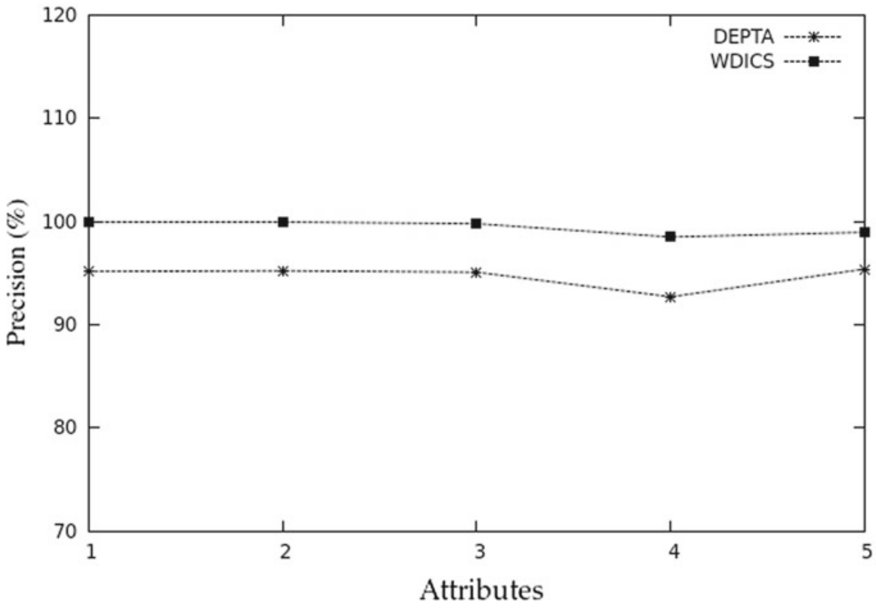


Fig. 2.3 Precision versus attributes

calculated based on the records extracted by our model, the records found by the search engine and the total available records on the Bluetooth website. For different attributes as shown in Table 2.2, the Recall and Precision are calculated and their comparisons are as shown in Figs. 2.3 and 2.4, respectively. It is observed that the Precision of WDICS increases by 4% and the Recall increases by 2% compared to DEPTA. Therefore, WDICS is more efficient than DEPTA because when an object is dissimilar to its neighbouring objects, DEPTA fails to identify all records correctly.

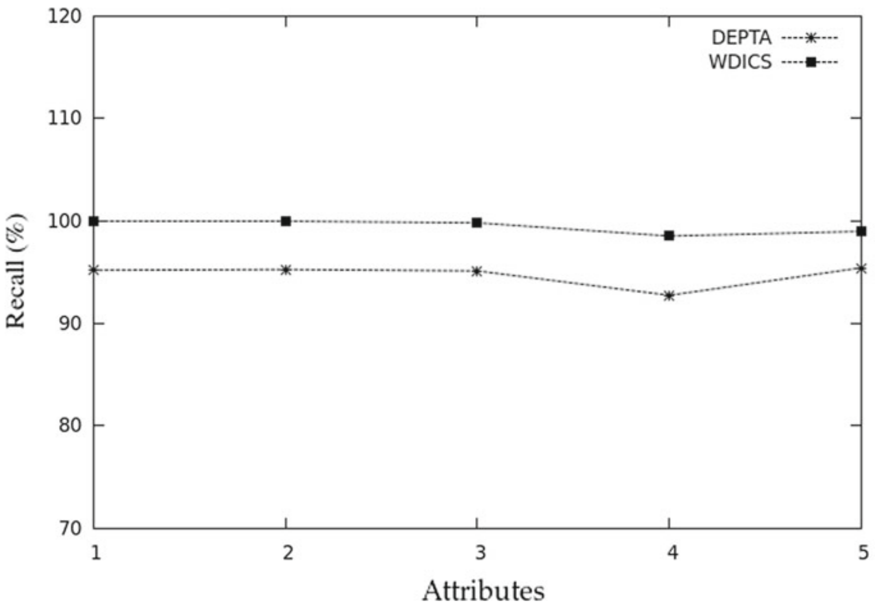


Fig. 2.4 Recall versus attributes

2.6 Summary

One of the major issues in Web Content Mining is the extraction of precise and meticulous information from the Web. In this chapter, we propose two similarity-based mechanisms; WDES to extract desired SERPs and store them in the local depository for offline browsing and WDICS to integrate the requested contents and enable the user to perform the intended analysis and extract the desired information. This results in faster data processing and effective offline browsing for saving time and resources. Our experimental results show that WDES and WDICS outperform DEPTA in terms of precision and recall. Further, different Web mining techniques such as classification and clustering can be associated with our approach to utilize the integrated results more efficiently.

References

1. Y. Zhai, B. Liu, Structured data extraction from the web based on partial tree alignment. J. IEEE TKDE **18**(12), 1614–1627 (2006)
2. G. Ananthanarayanan, S. Blagsvedt, K. Toyama, OWEB: a framework for offline web browsing, in *Fourth Latin America Web Congress, IEEE Computer Society* (2006), pp. 15–24
3. J. Myllymaki, Effective web data extraction with standards XML technologies, in *Proceedings of the 10th International Conference on World Wide Web* (2001), pp. 689–696

4. J. Yang, G. Shi, Y. Zheng, Q. Wang, Data extraction from deep web pages, in *IEEE International Conference on Computational Intelligence and Security* (2007), pp. 237–241
5. G.-S. Yin, G.-D. Guo, J.-J. Sun, A template-based method for theme information extraction from web pages, in *IEEE International Conference on Computer Application and System Modelling (ICCASM 2010)*, vol. 3 (2010), pp. 721–725
6. X. Liu, H. Li, D. Wu, J. Huang, W. Wang, L. Yu, Y. Wu, H. Xie, On web page extraction based on position of DIV, in *IEEE 4th ICCAE* (2010), pp. 144–147
7. N. Dalvi, P. Bohannon, F. Sha, Robust web extraction: an approach based on a probabilistic tree-edit model, in *Twenty-Eight ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (2009), pp. 335–348
8. R. Novotny, P. Vojtas, D. Maruscak, Information extraction from web pages, in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops* (2009), pp. 121–124
9. T. Nie, Z. Wang, Y. Kou, R. Zhang, Crawling result pages for data extraction based on URL classification, in *IEEE 7th Web Information Systems and Application Conference* (2010), pp. 79–84
10. N.K. Papadakis, D. Skoutas, K. Raftopoulos, STAVIES: a system for information extraction from unknown web data sources through automatic web wrapper generation using clustering techniques. J. IEEE TKDE **17**(12), 1638–1652 (2005)
11. C.-H. Chang, M. Ramzy, M.R. Girgis, A survey of web information extraction systems. J. IEEE TKDE **18**(10), 1411–1428 (2006)
12. N. Angkawattanawit, A. Rungsawang, Learnable crawling: an efficient approach to topic-specific web resource discovery, in *Proceedings of the 2nd International Symposium on Communications and Information Technology* (2002), pp. 97–114
13. K. Bharat, M.R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1998), pp. 104–111
14. C. Aggarwal, F. Al-Garawi, P. Yu, Intelligent crawling on the world wide web with arbitrary predicates, in *Proceedings of the 10th International World Wide Web Conference* (2001), pp. 96–105
15. S. Chakrabarti, K. Punera, M. Subramanyam, Accelerated focused crawling through online relevance feedback, in *WWW* (ACM, 2002)
16. M. Ehrig, A. Maedche, Ontology-focused crawling of web documents, in *Proceedings of the 2003 ACM Symposium on Applied Computing* (2003), pp. 1174–1178
17. K.C. Srikantaiah, M. Suraj, K.R. Venugopal, S.S. Iyengar, L.M. Patnaik, Similarity based web data extraction and integration system for web content mining, in *Proceedings of the 3rd International Conference on Advances in Communication, Network and Computing Technologies, CNC 2012, LNICST* (2012), pp. 269–274
18. Bluetooth SIG Website, https://www.bluetooth.org/tpg/listings.cfm