Information extraction from web tables

Conference Paper · December 2009

DOI: 10.1145/1806338.1806426 · Source: DBLP

4 authors:

Mahmood Shakir Hammoodi
University of Babylon
15 PUBLICATIONS **51** CITATIONS
SEE PROFILE

Hamidah Ibrahim
Universiti Putra Malaysia
230 PUBLICATIONS **1,380** CITATIONS
SEE PROFILE

Aida Mustapha
Universiti Tun Hussein Onn Malaysia
383 PUBLICATIONS **3,677** CITATIONS
SEE PROFILE

Lili N. Abdullah
Universiti Putra Malaysia
61 PUBLICATIONS **354** CITATIONS
SEE PROFILE

Information Extraction from Web Tables

Mahmoud Shaker[1], Hamidah Ibrahim[2], Aida Mustapha[3], Lili Nurliyana Abdullah[4]
Department of Computer Science
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia, 43400 Serdang, Malaysia
[1]Mah222254@yahoo.com, [2]hamidah@fsktm.upm.edu.my, [3]aida@fsktm.upm.edu.my,
[4]liyana@fsktm.upm.edu.my

ABSTRACT
Nowadays, many users use web search engines to find and gather information. User faces an increasing amount of various web pages information sources. The issue of correlating, integrating and presenting related information to users becomes important. When a user uses a search engine such as Yahoo and Google to seek a specific information, the results are not only information about the availability of the desired information, but also information about other pages on which the desired information is mentioned. Extracting information from the web pages also becomes very important because the massive and increasing amount of diverse web pages information sources in the Internet that are available to users, and the variety of web pages making the process of information extraction from web a challenging problem. This paper proposes an approach for extracting information from web tables based on standard classifications. The proposed approach consists of four main phases, namely: (i) pre-processing, (ii) extraction, (iii) classification, and (iv) simplification. The proposed approach is evaluated by conducting experiments on a number of web pages from the Nokia products domain, as to the best of our knowledge this is the only product that has complete and complex standard classifiers.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Information filtering, Search process, Selection process*.

General Terms
Standardization.

Key words
Information Extraction, Web Tables.

1. INTRODUCTION
At the present time, the Internet is general and many people use the Internet to find information. A variety of web pages and the frequently changing of information in web pages make searching and extracting information very difficult.

iiWAS2009, December 14–16, 2009, Kuala Lumpur, Malaysia.

When Internet users want to get information, they first visit search engines such as Yahoo and Google, and then visit all web sites suggested by the search engine. Many researchers such as [3], [4], [10], and [11] work on extraction of information from web pages in different domains (traveling, products, business intelligence) but these researches deal with limited web pages and the user still need to use the search engines such as Yahoo and Google to collect more information. Many of the web pages that the corporations used to announce their products (Internet shops) in tabular form consist of attributes, sub attributes, and values of sub attributes [1] [3]. Kostyantyn Shchekotykhin, *et al* [8] explained that much useful information is presented in tabular form on the web pages and Wolfgang Gatterbauer, *et al* [12] showed that extracting information from web tables is possible without any training and without reliance on heavy linguistic techniques tuned to the domain of interest in addition to the tables are interesting because they present information in a condensed, rather simple, and well structured way. Tables on the web pages are used for both (i) the genuine purposes that are presenting certain types of data to users which are formatted in rows and columns and (ii) helping construct the layout of a web page. Thus, tables are the richest sources of information on the web pages.

The rest of the paper is organized as follows. In section 2, the previous works related to this research are reported. In section 3, we present the structure of the Standard Classifications (SC) and the structure of Web Pages (WP). In section 4, we present an overview of the proposed approach. In Section 5, we present the experimental results. Conclusion is presented in the final section, 6.

2. RELATED WORKS
This section presents a review of previous approaches that have been proposed for extracting information from web pages which can be organized based on the type of technique used by each approach to extract information, namely: (i) Semantic-based, (ii) Structure-based, (iii) XML-based, (iv) Ontology-based, and (v) Clustering-based. The details of each approach are discussed below.

Semantic-based: With the advent of the Internet, more information is available electronically, and the information on the Internet is generated in textual form which differs from the web page to another in semantics. Semantics generally deal with the relationships between signs and concepts (mental signs). Different kinds of semantics are Lexical Semantics, Statistical Semantics, Structural Semantics, and Prototype Semantics. Srinivas Vadrevu, *et al* [10] have focused on information extraction from web pages

using presentation regularities and domain knowledge. They argued that there is a need to divide a web page into information blocks or several segments before organizing the content into hierarchical groups and during this process (partition a web page) some of the attribute labels of values may be missing.

Structure-based: The structure based approaches employ assumptions about the general structure of tables (i.e., <TABLE> tags) on the web pages [12]. Wolfgang Gatterbauer, *et al* [12] propose an approach for extracting information from web tables. Their approach analyzes any given web page for the existence of tabular data, recognizes the relations as implied by their spatial arrangement, extracts a number of *n*-tuples together with hierarchical information about relations between their entries, and saves them in structured data format. The approach employs a model of visual representation of web pages rendered by a web browser. In addition, loading and rendering a web page happens in *n* times, where *n* is the number of element nodes as measure of a web page's size. However, this is the obstacle of their approach. Jeong-Woo Son *et al* [5] have proposed an approach to discriminate web tables using a composite kernel which combines a parse tree kernel and a linear kernel. They proposed three kinds of features to capture both kinds of web table information which is composed of structural and content ones. First, the parse tree is adopted to reflect the structural information. Second, the content type features are adopted to capture the content information. Finally, they combined both kinds of information using a composite kernel. The main obstacle of their approach comes from the difficulty of generating relevant features for the discrimination.

XML-based: There are several challenges in extracting information from a semi-structured web page such as the lack of a schema, ill formatting, high update frequency, and semantic heterogeneity of the information. In order to overcome these challenges, some researchers have proposed approaches for transforming the page into a format called Extensible Mark-up Language (XML) [9]. Man I. Lam, *et al* [9], proposed a system which used different methodologies to extract the information. The extraction task is only individual page based. It means that all the fields for the same record are supposed to be contained in the same page. However, in many other situations, the fields may be located in different relevant pages, such as several linked web pages.

Ontology-based: Ontology is a branch of philosophy and structures of objects, properties, events, processes and relations in every area of reality. Horacio Saggion, *et al* [4], proposed the MUSING project (Multi-industry, Semantic-based next generation business intelligence). The MUSING project needs to cover many semantic categories including locations, organizations and specific business events to help companies that want to take their business overseas and concerned in knowing the best place to exploit.

Clustering-based: Cluster analysis has been playing an important role in solving many problems in medicine, psychology, biology, sociology, pattern recognition, and image processing. Clustering algorithms attempt to assess the interaction among patterns by organizing patterns into clusters such that patterns within a cluster are more similar to each other than are patterns belonging to different clusters [2]. Fatima Ashraf, *et al* [2] have employed clustering techniques for automatic information extraction from HTML documents containing HTML data. They proposed a system which is called ClusTex. If the tokens of one kind differ from each other in format, then this leads to an incorrect clustering of some tokens.

The web tables provide more organized information, summarized information, and conciseness in expressing knowledge [5]. Therefore, focus is given more on the structure-based which is the main focus of this paper.

3. THE STRUCTURE OF THE STANDARD CLASSIFICATIONS (SC) AND WEB PAGES (WP)

The structure of the standard classifications consists of attributes, sub attributes, and groups of the sub attributes. The following explains the structure of the standard classifications [3] [6]:

(i) Attribute describes the properties of a product. Each product usually has a description of its properties and various aspects of its use. For example the attributes which are used for describing the properties of Nokia product are Size, Display, Memory, Data, etc.

(ii) Sub attribute describes the properties of an attribute. For example: Width, Height, Weight, etc describe the attribute Size.

(iii) Group of sub attributes, the sub attributes that belong to the same attribute are grouped together in a group. For example, Width, Height, and Weight that belong to the attribute Size are grouped in the same group.

We use Attr (SC), Sub_Attr (SC), and G_Sub (SC) to denote the attributes of SC, the sub attributes of SC, and group of sub attributes, respectively.

Most web pages have similar structure as the SC that are attributes, sub attributes, and groups of sub attributes with additional element, value which describes the value of a sub attribute. For example, class32 and 123 kbps are the values of GPRS which is one of the sub attributes that describes the attribute Data. There are web pages with simpler structures that contain only attributes and values of attributes. The symbol Attr (WP), Sub_Attr (WP), and G_Sub (WP) denote the attributes of WP, the sub attributes of WP, and group of the sub attributes, respectively.

4. AN OVERVIEW OF THE PROPOSED APPROACH

The proposed approach consists of four main phases, namely: (i) pre-processing, (ii) extraction, (iii) classification, and (iv) simplification. Figure 1 gives an overview of the proposed approach.

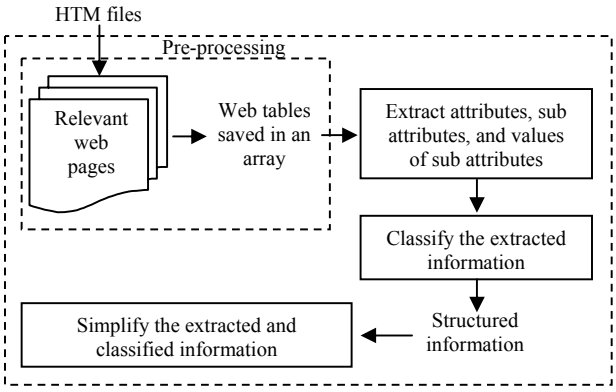


Figure 1. The process of extracting, classifying, and simplifying information from web tables

4.1 Pre-processing Phase

Not all of the web pages that are received from search engine are related to user's desire. Therefore, the relevant web page is determined by analyzing the title of a web page. The proposed approach checks the title of each web page by comparing the tokens which are found between the tag <TITLE> and </TITLE> with a table consisting of a list of product names. Figure 2 illustrates an example of source code with title of a web page being highlighted.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//E
<!-- saved from url=(0047)http://www.esato.com/phones/index.php/phone=
<HTML>
<HEAD>
<TITLE>Nokia 7600 - specifications and reviews</TITLE>
<META http-equiv=Content-Type content="text/html; charset=ISO-8859-1"
<META content="Nokia 7600 specifications and reviews" name=keywords
<META
content="Nokia 7600 specifications and reviews and pictures of the phone r
name=description><LINK href="/" rel=top><LINK title="Esato.com RSS"
href="http://www.esato.com/rss/" type=application/rss+xml rel=alternate><
href="/help/downloadhelp.php" rel=help><LINK href="/about/sitemap.php"
rel=contents><!-- 2008-09-24 19:25:29 -->
<SCRIPT language=javascript src="21_files/script.js"
type=text/javascript></SCRIPT>

<SCRIPT language=javascript type=text/javascript><!--function rate(pid){
"/phones/comparesize.php?id=" + id + "&id2=" + id2;width = w1 + w2;if(h1>
param);}function ovcomp(id1, id2, w, h){    flt = document.getElementById
= "url('/gfx/phonecomparebackground.png')";          var iHTML = 'Relati
src="http://www.esato.com/phones/comparesize_image.php?id=' + id1 + '">
'"></tr></table><hr/>'          flt.innerHTML = iHTML;}function hwcom
```

Figure 2. Example of source code (title of a web page)

Then, the tokens which are found between the tag <TABLE> and </TABLE> are saved in an array for matching them with SC. The tag <TR> denotes the row of <TABLE>, and the tag <TD> denotes the field of <TR>. If there is more than one tag <TD> then the tokens are saved and prefixed with the symbol "-" which denotes a sub attribute (WP), and symbol ":" which denotes the value of a sub attribute (WP). If there is only one <TD> in one of <TR> then the token is saved with prefix "*" which denotes an attribute (WP). Figure 3 (a) illustrates an example of a source code (WP) with the tags <TABLE>, <TR>, and <TD>. Figure 3 (b) illustrates the sub attributes and values of the sub attributes found in Figure 3 (a) saved with the symbols "-" and ":" in an array. For example, the sub attribute Brand saved with the symbol "-" which denotes a sub attribute (WP) and the value Nokia with the symbol ":" which denotes the value of a sub attribute (WP).

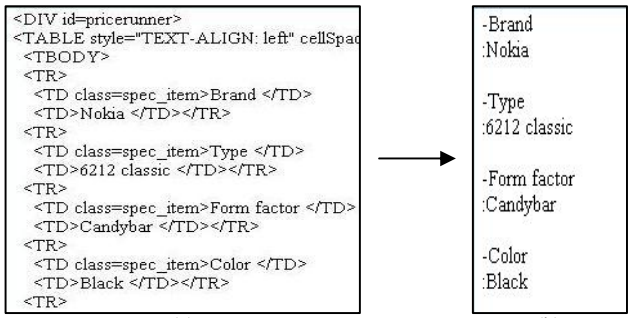


Figure 3. Example of a source code (WP) with the tags <TABLE>, <TR>, and <TD> consisting of sub attributes (WP) and values of sub attributes (WP) saved in an array

4.2 Extraction Phase

The extraction phase consists of three main rules for extracting relevant information from web tables, as discuss below.

4.2.1 Rule for Extracting Attributes, Sub Attributes, and Values of the Sub Attributes

The tokens which are saved in an array are then matched with Attr (SC). If there is a match then the Attr (WP), Sub_Attr (WP), and value of Sub_Atrr (WP) are extracted. The rule applied is as follows.

Rule 1:
If Attr (WP) = Attr (SC) then
Extract Attr (WP), Sub_Attr (WP), and value of Sub_Attr (WP)

Figure 4 illustrates an example of a web page that is used to announce Nokia product.

Figure 4. Example of attributes, sub attributes, and values of the sub attributes

The attribute DATA (WP) matched with the attribute DATA (SC) and based on Rule 1, the proposed approach extracts the attribute, the sub attributes of DATA that are GPRS, HSCSD, EDGE, 3G, WLAN, Bluetooth, Infrared Port, and USB that describe the extracted attribute, and values of each of the sub attributes.

There are cases where no match is found between Attr (WP) and Attr (SC). For such cases, another rule is applied as discuss below.

4.2.2 Rules for Extracting Sub Attribute and Value of the Sub Attribute

There are two cases to be considered, namely: (i) matching the Attr (WP) with the Sub_Attr (SC) and (ii) matching G_Sub (WP) with each G_Sub (SC).

4.2.2.1 Match Attr (WP) with Sub_Attr (SC)

In some of the web pages, the sub attribute appears as attribute. Therefore, the Attr (WP) is match against the Sub_Attr (SC). If there is a match then the Attr (WP) is extracted as a sub attribute together with its value. The rule applied is as follows.

Rule 2:
If Attr (WP) = Sub_Attr (SC) then
Extract Attr (WP) as a sub attribute and value of the sub attribute

The structure of the web page in Figure 5 consists of attributes and values of the attributes. The sub attributes appear as attributes. For example, the attributes Height and Width which are actually sub attributes in the SC that belong to the attribute SIZE appear as attributes in Figure 5.

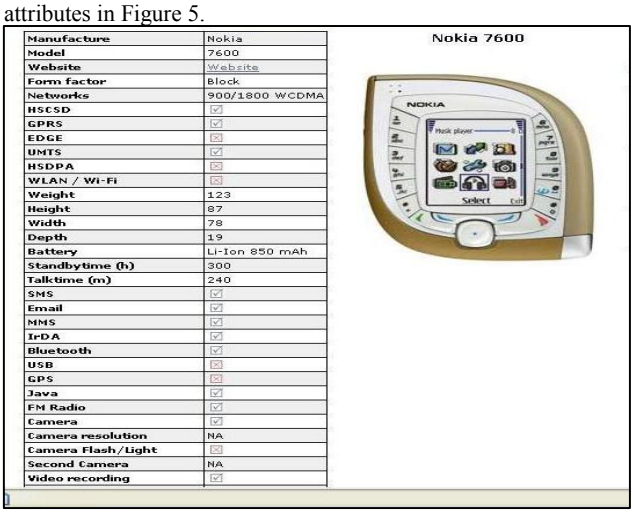


Figure 5. Example of sub attributes appear as attributes

The attribute Width (WP) matched with the sub attribute Width (SC) which describes the attribute SIZE. Therefore, the attribute Width (WP) is extracted as sub attribute.

4.2.2.2 Match G_Sub (WP) with Each G_Sub (SC)

Sometimes an attribute (WP) appears in different names which are not found in the standard classifications (SC), therefore the G_Sub (WP) that describes the Attr (WP) which appears in different name is match with each G_Sub (SC). The number of sub attributes from each G_Sub (SC) that is matched with G_Sub (WP) is saved in an array. The G_Sub (SC) with the maximum number of matched sub attributes is selected and the Attr (WP), G_Sub (WP), and values of the sub attributes are then extracted as shown in Figure 6. The rule applied is as follows.

Rule 3:
If G_Sub (WP) ⊆ G_Sub (SC) then
Maximum_array ← the number of sub attributes (SC) that matched
Select the Attr (WP) and G_Sub (WP), and values of the sub attributes with the maximum number of matched sub attributes from *Maximum_array*

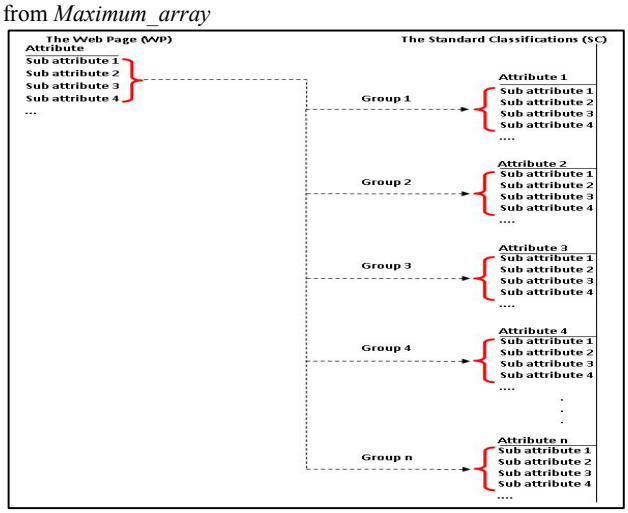


Figure 6. A group of the sub attributes (WP) is matched with each group of the sub attributes (SC)

Figure 7 illustrates another example of a web page with similar structure as the web page in Figure 4. If we compare the attributes of Figure 4 and Figure 7, it is found that the attributes have different names and the same attribute may contain different kinds of sub attributes. For example the attribute Memory in Figure 4 consists of the sub attributes Phonebook, Call records, and Card slot while in Figure 7, the same attribute consists of the sub attributes Internal memory, External memory, Memory slots, and Storage types. In addition, the attribute DATA in Figure 4 appears as Connectivity in Figure 7 that consists of the same sub attributes such as Bluetooth and USB which appear in Figure 4.

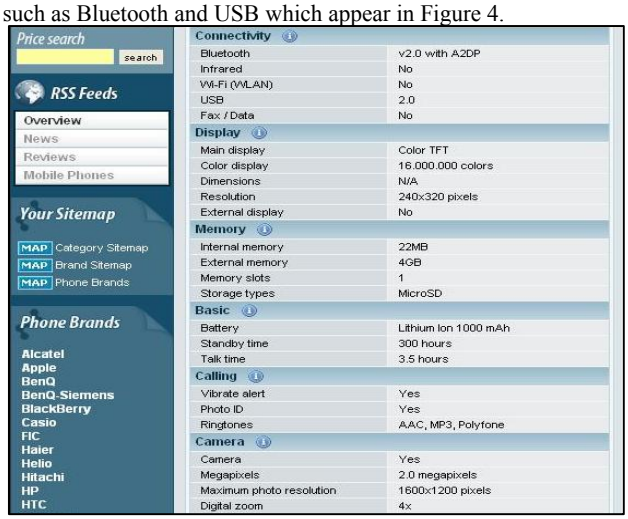


Figure 7. Example of attributes, sub attributes, and values of the sub attributes

The attribute Connectivity (WP) in Figure 7 is not found in the SC, by applying Rule 3, the group of the sub attributes that

describes the attribute Connectivity (WP) is matched with each group of the sub attributes (SC). Figure 8 presents the rules applied for the web page given in Figure 7.

Attr (WP)	Sub_Attr (WP)	Rules Applied	Description
Connectivity	Bluetooth	Rule 3	Applied based on the sub-attributes
	Infrared	Rule 3	
	Wi-Fi(WLAN)	Rule 3	
	USB	Rule 3	
	Fax / Data	Rule 3	
Display	Main display	Rule 1	Applied based on the attribute Display
	Color display		
	Dimensions		
	Resolution		
	External display		
Memory	Internal memory	Rule 1	Applied based on the attribute Memory
	External memory		
	Memory slots		
	Storage types		
Basic	Battery	Rule 3	Applied based on the sub-attributes
	Standby time	Rule 3	
	Talk time	Rule 3	
Calling	Vibrate alert	Rule 3	Applied based on the sub-attributes
	Photo ID	Rule 3	
	Ringtones	Rule 3	
Camera	Camera	Rule 3	Applied based on the sub-attributes
	Megapixels	Rule 3	
	Maximum photo resolution	Rule 3	
	Digital zoom	Rule 3	

Figure 8. The rules applied for the web page shown in Figure 7

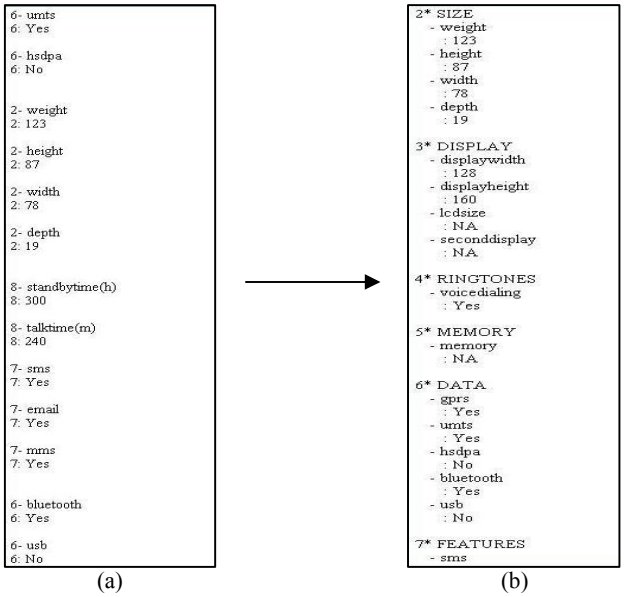
4.3 Classification Phase

The extracted information is classified by identifying the index number of Attr (SC) that is matched and the extracted attributes and sub attributes are then grouped based on the index number. The Attr (WP), Sub_Attr (WP), and value of Sub_Attr (WP) are saved in a text file with the index of Attr (SC) that matched. Figure 9 illustrates the example of the attributes that are saved in database with the index number *Index_no*.

Attribute	Index_no
General	1
Size	2
Display	3
Ringtones	4
Memory	5
Data	6
Features	7
Battery	8

Figure 9. Attr (SC) saved in database, *Index_no* denotes the index of Attr (SC)

Figure 10 (a) illustrates the example of the sub attributes and values of the sub attributes, where each line begins with the index of Attr (SC) that is matched. For example, the sub attribute weight is saved with the index of the attribute SIZE.



(a) (b)

Figure 10. Attr (WP) in a text file with index number of Attr (SC)

The matched attributes and sub attributes are then grouped based on the index number. For example, the lines with the index 6 are grouped together as attribute DATA, as shown in Figure 10 (b) which illustrates the example of the extracted attributes and sub attributes after grouping them based on the index number.

In Figure 10 (b), the symbol "*" denotes Attr (WP), the symbol "-" denotes Sub_Attr (WP), and the lines without the symbols "*" and "-" represent the value of Sub_Attr (WP). The extracted information is saved in a text file. Figure 10 (b) illustrates an example of a text file.

Next, the name of the text file, path of the text file, name of product, number of matched sub-attribute value pairs extracted (WP), and number of unmatched sub-attribute value pairs extracted (WP) are saved in a table (*Structured Information*). Figure 11 illustrates an example of the extracted results.

	Name of Text	Path of Text	Product	No of correct values extracted	No of incorrect values extracted	Total number of possible correct values
▶	Text 1	F:\my project2\WP in Text after classify\1.txt	Nokia N79	52	2	54
	Text 2	F:\my project2\WP in Text after classify\2.txt	Nokia 7600	53	1	54
	Text 3	F:\my project2\WP in Text after classify\3.txt	Nokia 6212 classic	49	0	49
	Text 4	F:\my project2\WP in Text after classify\4.txt	Nokia 6600 fold	28	0	28
	Text 5	F:\my project2\WP in Text after classify\5.txt	Nokia 7310 Supernova	27	0	27
	Text 6	F:\my project2\WP in Text after classify\6.txt	nokia 7600	14	10	24
	Text 7	F:\my project2\WP in Text after classify\7.txt	Nokia 5800	32	0	32
*						

Figure 11. Example of the extracted results

4.4 Simplification Phase

The relevant information that is extracted is then analyzed to identify the attributes and sub attributes that belong to the same product which are extracted repetitively. So that, these repetitive attributes and sub attributes are removed. There are two main steps for analyzing the extracted relevant information.

4.4.1 Group the Records with the Same Name of a Product in a Table

The records are grouped in *Structured Information* based on the name of the product. Those records with the same product name are saved in the same table (*Similar Table*). For example, there are two text files in Figure 11 that are Text 2 consisting of 53 extracted sub attributes and Text 6 consisting of 14 extracted sub attributes for the same product Nokia 7600. Text 2 and Text 6 are then saved in the same table.

4.4.2 Compare the Extracted Sub Attributes that Belong to the Same Product

The extracted sub attributes that belong to the same product are compared and the attributes and sub attributes that are duplicates are removed. For example, refer to Text 2 and Text 6 shown in Figure 11. The sub attributes of Text 2 and Text 6 are compared. Text 2 consists of 53 extracted sub attributes while Text 6 consists of 14 extracted sub attributes which are found to be part of the extracted attributes of Text 2. Therefore, Text 6 is removed. Figure 12 illustrates example of the extracted information.

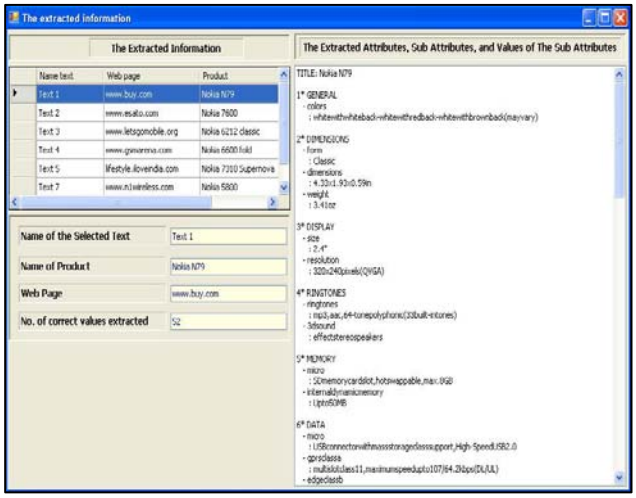


Figure 12. Example of the extracted information

5. EXPERIMENTAL RESULTS

In this section, we present details of the experiments followed by discussion and comparison with those reported in the literature. To evaluate our approach, the following web sites were selected that are www.buy.com "Cell Phones and Services" which is used by Fatima Ashraf, *et al* [2], www.gsmarena.com, www.esato.com, www.letsgomobile.org, and lifestyle.iloveindia.com which are used to announce the products of Nokia mobile phone.

5.1 Evaluation

The parameters used to evaluate our approach are precision, recall, and the geometrical average of these two, the *F* value. The *F measure* can be defined to have a metric that can be used to compare various IE systems by only one value [7]. Researchers in the IE field commonly report their result by using these metrics.

Precision (P) = C / (C+I)
Recall (R) = C / T

where
C: Number of correct sub-attribute value pairs extracted,
I: Number of incorrect sub-attribute value pairs extracted, and
T: Total number of possible correct sub-attribute value pairs.

$$f = \frac{\left(\beta^2 + 1\right) * P * R}{\beta^2 * P + R}$$

where β^2 is the weight of R over P, a value of $\beta^2 = 1$ means that recall and precision are weighted equally. Fatima Ashraf, *et al* [2] reported the F value where β^2 is taken to be 1.

5.2 Experiments and Results

Fatima Ashraf, *et al* [2] tested their approach on www.buy.com "Cell Phones and Services", and they reported P = 94.55%, R = 100%, and F = 97.19%. This web page contains the Manufacturer, the Cell Phone Model, and the Price. In their work, if the tokens of one kind differ from each other in format, then this would lead to an incorrect clustering of some tokens. Our approach extracts the attributes which are Size, Display, Ringtones, Memory, Data, Features, and Battery from the web site www.buy.com besides the sub attributes that describe the attributes and values of the sub attributes. While the same attributes, sub attributes, and values of the sub attributes in addition to the attribute General are extracted from the web sites www.gsmarena.com, www.esato.com, www.letsgomobile.org, and lifestyle.iloveindia.com. We reported P = 99.07%, R = 99.07%, and F = 99.07% as shown in Table 1.

Table 1. Extraction results from our approach compared to Fatima Ashraf, *et al* [2]

The approaches	Precision	Recall	F
The proposed approach	99.07%	99.07%	99.07%
Previous approach [2]	94.55%	100.00%	97.19%

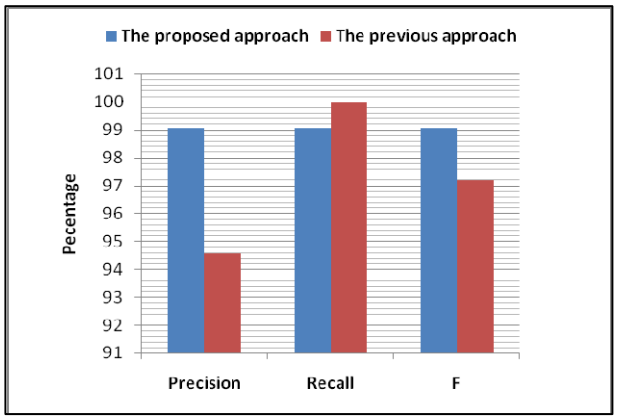


Figure 13. Extraction results from our approach compared to Fatima Ashraf, *et al* [2]

Figure 13 illustrates the increment in precision and *F* measure that is achieved in our approach, and decrement in recall. The ratio of increment in precision is 4.52%, the ratio of decrement in recall is 0.93%, and the ratio of increment in *F* is 1.88%. Katharina Kaiser and Silvia Miksch [7] explained that if a system optimized for high precision the feasibility of not detecting all relevant information improves while if recall is optimized it is possible that the system classifies irrelevant information as relevant.

Nonetheless, due to the lack of test data sets, it is difficult to compare our approach with the work of Wolfgang Gatterbauer, *et al* [12] and Jeong-Woo Son, *et al* [5] which we have referenced in related works proposed approaches for extracting information from web tables. On the other hand, their approaches did not handle the attributes which appear under different names but refer to the same entity (i.e., synonyms). This issue has been addressed by our approach.

6. CONCLUSION

In this paper, we proposed an approach for extracting relevant information from various web pages. Experiments demonstrated that our approach extracts the attributes besides the sub attributes that describe the extracted attributes and values of the sub attributes from various web pages. Besides, the proposed approach is able to extract the attributes that appear under different names but refer to the same entity (i.e., synonyms).

There are a number of suggestions to extend this work. One direction is to link the presented work to various search engines such as Msn, Yahoo, Google, *etc* to search relevant information based on the user's queries for extracting information from various web pages obtained from different search engines. Besides, a high ranking for a specific keywords in one search engine does not automatically mean that the obtained web pages will rank highly for the same keywords in another search engine. Another direction is to add an approach for parsing the web pages which are not based on the English language.

7. REFERENCES

[1] Fatima Ashraf and Reda Alhajj, 2007. ClusTex: Information Extraction from HTML Pages. In Proceedings of the 21st. International Conference on Advanced Information Networking and Applications Workshops (AINAW'O7). 1: 355-360. DOI= 10.1109/AINAW.2007.119

[2] Fatima Ashraf, Tansel Ozyer, and Reda Alhajj, 2008. Employing Clustering Techniques for Automatic Information Extraction from HTML Documents. Journal of IEEE Transactions on Systems. 38: 660-673. DOI= 10.1109/TSMCC.2008.923882

[3] Guntis Arnicans and Girts Karnitis, 2006. Intelligent Integration of Information from Semi-Structured Web Data Sources on the Base of Ontology and Meta-Models. In Proceedings of the 7th. International Baltic Conference. 177-186. DOI= 10.1109/DBIS.2006.1678494

[4] Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva, 2007. Ontology-based Information Extraction for Business Intelligence. In Proceedings of the 6th. International Semantic Web Conference and the 2nd. Asian Semantic Web Conference. United Kingdom. 843-856. DOI= 10.1007/978-3-540-76298-0_61

[5] Jeong-Woo Son, Jae-An Lee, Seong-Bae Park, Hyun-Je Song, Sang-Jo Lee, and Se-Young Park, 2008. Discriminating Meaningful Web Tables from Decorative Tables using Composite Kernel. In Proceedings of ACM International Conference on Web Intelligence and Intelligent Agent Technology. 1:368-371. DOI: 10.1109/WIIAT.2008.241

[6] Jyotirmaya Nanda, Timothy W. Simpson, Soundar R. T. Kumara, and Steven B. Shooter, 2006. A Methodology for Product Family Ontology Development using Formal Concept Analysis and Web Ontology Language. Journal of Computing and Information Science in Engineering. 6:1-11. DOI= 10.1115/1.2190237

[7] Katharina Kaiser and Silvia Miksch, 2007. Modeling Treatment Processes using Information Extraction. In: Advanced Computational Intelligence Paradigms in Healthcare – 1, 84:189-224. Springer Berlin, Heidelberg. DOI= 10.1007/978-3-540-47527-9

[8] Kostyantyn Shchekotykhin, Dietmar Jannach, and Gerhard Friedrich, 2007. Clustering Web Documents with Tables for Information Extraction. In Proceedings of the 4th. International Conference on Knowledge Capture, Canada, 169-170. http://doi.acm.org/10.1145/1298406.1298438

[9] Man I. Lam, Zhiguo Gong, and Maybin Muyeba, 2008. A Method for Web Information Extraction. In Proceedings of 10th. Asia-Pacific Web Conference. APWeb. Shenyang. China. 4976: 383-394. DOI= 10.1007/978-3-540-78849-2

[10] Srinivas Vadrevu, Fatih Gelgi, and Hasan Davulcu, 2007. Information Extraction from Web Pages using Presentation Regularities and Domain Knowledge. Journal of World Wide Web, Springer Netherlands. Arizona State University. USA, 10: 157-179. DOI= 10.1007/s11280-007-0021-1

[11] Sung Won Jung, Kyung Hee Sung, Tae Won Park, and Hyuk Chul Kwon, 2001. Intelligent Integration of Information on the Internet for Travelers on Demand. In Proceedings of ISIE IEEE International Symposium. Pusin, Korea, 338-342. DOI= 10.1109/ISIE.2001.931810

[12] Wolfgang Gatterbauer, Paul Bohunsky, Marcus Herzog, Bernhard Krupl, and Bernhard Pollak, 2007. Towards Domain-independent Information Extraction from Web Tables. In Proceedings of the 16th. International Conference on World Wide Web. Canada, 71-80. http://doi.acm.org/10.1145/1242572.1242583