5

# A Framework for Extracting Information from Semi-Structured Web Data Sources

Mahmoud Shaker[1], Hamidah Ibrahim[2],
Aida Mustapha[3] and Lili Nurliyana Abdullah[4]
*Department of Computer Science*
*Faculty of Computer Science and Information Technology*
*Universiti Putra Malaysia, 43400 Serdang,*
*Malaysia*

## 1. Introduction

Nowadays, many users use web search engines to find and gather information. User faces an increasing amount of various semi-structured 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. The number of selected pages is enormous. Therefore, the performance capabilities, the overlap among results for the same queries and limitations of web search engines are an important and large area of research. Extracting information from the web data sources also becomes very important because the massive and increasing amount of diverse semi-structured 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 chapter presents a framework for extracting and classifying semi-structured web data sources. The framework is able to extract relevant information from different web data sources, and classify the extracted information based on a given standard classification. In this chapter, we focus on the Nokia products, as to the best of our knowledge this is the only product that has complete and complex standard classifiers. 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 data sources make searching and extracting information very difficult. When Internet users want to get information about Nokia products for example, 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 Guntis Arnicans and Girts Karnitis 2006; Sung Won Jung *et al* 2001; Srinivas Vadrevu *et al* 2007; and Horacio Saggion *et al* 2008 work on extraction of information from web data sources in different domains (traveling, products, business intelligence) but these researches deal with limited web data sources and users still need to use the search engines such as Yahoo and Google to collect more information. We proposed a framework for extracting information from different web data sources. The components of the proposed framework include *Query*

Source: Convergence and Hybrid Information Technologies, Book edited by: Marius Crisan,
ISBN 978-953-307-068-1, pp. 426, March 2010, INTECH, Croatia, downloaded from SCIYO.COM

*Interface* (QI) which is used for accepting user's queries and searching web pages based on the user's queries through search engine, *Information Extraction* (IE) which is used for extracting and classifying the web pages obtained from QI and converting the extracted and classified information into text form, and *Relevant Information Analyzer* (RIA) which is used for determining the relevant information extracted from Information Extraction (IE). The rest of the chapter is organized as follows. In section 2, we explain the concepts related to a typical Information Extraction (IE). In section 3, the previous works related to this research are reported. In section 4, we present the proposed framework. Conclusion is presented in the final section 5.

## 2. Concepts of Information Extraction (IE)

Information Extraction (IE) is originally the task of locating specific information from a natural language document and is a particular useful sub-area of Natural Language Processing (NLP). The dramatic growth in the number and size of on-line textual information sources has led to an increasing research interest in the information extraction problem (Line Eikvil 1999). Information Extraction is a form of shallow document processing that involves populating a database with values automatically extracted from documents. Over the past decade, researchers have developed a rich family of generic Information Extraction techniques that are suitable for a wide variety of sources from rigidly formatted documents such as HTML generated automatically from a template to natural-language documents (Nicholas Kushmerick 2003). Information Extraction promises to be a sizeable augmentation to the search engines available today, and it can extract precisely the information that the user wants from this set of documents, and provide the user with exactly the information that is required without the level of involvement that this task requires currently (Chia-Hui Chang *et al* 2006). Information Extraction is to discover relevant information without any training (Wolfgang Gatterbauer *et al* 2007). Information Extraction is the identification or pre-processing, consequent or concurrent classification, and structuring into semantic classes making the information more suitable for information processing tasks (Rik De Busser 2006). Information Extraction fills the fields in a table by automatically extracting sub-sequences of human readable text. Sub-sequences are the useful pieces of information in the documents which are taken as input to produce fixed format unambiguous data as output (Line Eikvil 1999; Chia-Hui Chang *et al* 2006). Figure 1 illustrates the taxonomy of Information Extraction which consists of different type of data as input and the approaches that have been proposed for extracting information from semi-structured data. The web tables provide more organized information, summarized information, and conciseness in expressing knowledge (Jeong-Woo Son et al 2008). Therefore, focus is given more on the structure-based which is the main focus of this chapter.

We can differentiate the various Information Extraction approaches by the type of data that are used as origin, namely: (i) structured data, (ii) semi-structured data, and (iii) unstructured data (Katharina Kaiser and Silvia Miksch 2007).

a. Structured Data: Structured data is a meaning of the particular data is assigned as well as it contains sufficient structure to allow unambiguous parsing and recognition of information. Thus, extracting relevant information and the assignment of a meaning can be eased (Katharina Kaiser and Silvia Miksch 2007) as well as quite simple techniques are sufficient for extracting information from structured data provided that the format is known (Line Eikvil 1999).

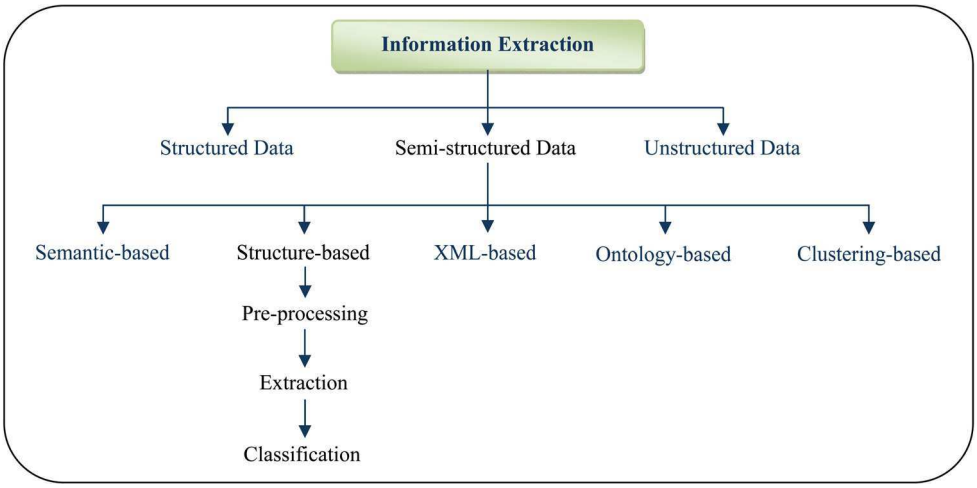


Fig. 1. The taxonomy of Information Extraction

b. Semi-Structured Data: Semi-structure means that it lacks of formatting structured. In semi-structured data, there is no separation between the data and the schema, and the amount of structure used depends on the purpose. No semantic is applied to semi-structured data and analysis of words or sentences is required for extracting the relevant information. XML and HTML pages contain semi-structured data, but HTML is rather more human-oriented or presentation-oriented (Katharina Kaiser and Silvia Miksch 2007). Semi-structured data refers to data with some of the following characteristics (Man I. Lam and Zhiguo Gong 2005):
   - The schema may be implicit in the data and it is not given in advance.
   - The schema may be changing frequently with respect to the size of the data. Therefore, the schema is relatively large.
   - The schema is attributive rather than prescriptive, i.e. it describes the current state of the data.
   - The values of the same attribute may be of differing types and the data is not typed powerfully.
   - The data transfer format may be portable.
   - It can represent the information of some data sources that cannot be constrained by schema.
   - It provides a flexible format for data exchange between different types of databases.

   Information Extraction from semi-structured data such as web pages which contain an enormous quantity of information which is usually formatted for human users is a useful yet complex task and provides a special challenge (Chia-Hui Chang *et al* 2006).

c. Unstructured Data: Unstructured data can be, for example, plain text. It does not imply that the data is structurally incoherent (in that case it would simply be nonsense), but rather that its information is encoded in such a way that makes it difficult for computers to immediately interpret it (Rik De Busser 2006). Linguistic knowledge is required to extract information from unstructured data as well as Natural Language Processing (NLP) techniques are deployed to design rules for locating specific pieces of

information or facts in the unstructured data (e.g., text) and using these facts to fill a database (Line Eikvil 1999; Katharina Kaiser and Silvia Miksch 2007).

Information Extraction is performed on unstructured data, semi-structured data, and structured data. Typically, techniques from Natural Language Processing (NLP) are often used for unstructured data and tend to be slow and this can be a problem as the volume of document collections on semi-structured data such as web pages can be large and the extraction is often expected to be performed on the fly. Therefore, NLP techniques are however not well suited for structured and semi-structured data as these techniques require full grammatical sentences (Line Eikvil 1999). The web pages provide a large and growing amount of information stores, which can be reached by manual browsing using a search engine. When a user does a keyword search on a search engine, a large number of web pages may result that might be very time consuming to check through. These web pages are often isolated with no effective connections between them, and respective access service live independently. This scenario motivates the need for information extraction from the web pages which can extract and group information from independent sources (Chia-Hui Chang *et al* 2006). Kostyantyn Shchekotykhin *et al* (2007) explained that much useful information is presented in tabular form on the web pages and Wolfgang Gatterbauer *et al* (2007) showed that extracting information from web tables is possible 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. David Buttler *et al* (2001) observed in their tests of 50 web sites with over 2000 web pages that the tag <TABLE> is used as object separator (18% of time) more than the other tags such as tag <P> 10% of time, tag <li> 8% of time, tag <hr> 6% of time, tag <ul> 2% of time, tag <DIV> 2% of time, and tag <a> 2% of time. Therefore, the relevant information in a web page that the user needs which must be extracted by IE are found between the tag <TABLE> and </TABLE> (Guntis Arnicans and Girts Karnitis 2006; Fatima Ashraf *et al* 2008). Each table is formatted in rows and columns, whereas it is distinguished in head and body according to meaning. In HTML documents, the tags such as <TABLE>, <TR>, and <TD> are reserved for table structure (Sung-won Jung *et al* 2001). Information Extraction systems do not attempt to understand the text in the input documents but they analyze those portions of each document that contain relevant information. Relevance is determined by predefined domain guidelines which specify what types of information that the system is expected to find (Line Eikvil 1999). Therefore, IE application needs lexicons to start with (e.g. attributes which specify the important information that the user needs to know for identification and utilization of the described objects). Thus, the acquisition and enrichment of lexical or semantic resources is a critical issue in Information Extraction. Standard Classification Scheme is used to identify the entities that have a form of existence in a given domain and specify their essential properties and it is a characterization vocabulary. Information Extraction techniques are then used to locate these entities from the web pages to be presented to the user (B. Chandrasokaran *et al* 1999; Stefan Clepce *et al* 2007).

## 3. Related works

The previous approaches are organized based on the type of technique used by each approach to extract information i.e. Semantic-based, Structure-based, XML-based, Ontology-based, and 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 deals 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* (2007) 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 (Wolfgang Gatterbauer *et al* 2007; Jeong-Woo Son *et al* 2008). Wolfgang Gatterbauer *et al* (2007) have proposed an approach for extracting information from web tables. Their approach analyzes any given web page for the existence of tabular data, recognizes 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 task of extracting web tables is formulated as the task of (i) finding all frames for a given web page, (ii) discerning those which adhere to the definition of tables where a 2-D grid is semantically significant from lists and other frames intended for non-relational layout purposes, (iii) transferring the content into a topological grid description in which logical cells are flush with neighboring cells and their spatial relations are explicit. Jeong-Woo Son *et al* (2008) 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) (Man I. Lam and Zhiguo Gong 2005). Man I. Lam and Zhiguo Gong (2005) 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* (2008) 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 (Fatima Ashraf *et al* 2008). Fatima Ashraf *et al* (2008) have employed clustering techniques for automatic information extraction from HTML documents containing HTML data. They proposed a system which is called ClusTex. They extend the work in Fatima Ashraf and Reda Alhajj (2007) by testing their proposed system in different domains such as Cell phone sales and Marathon schedule. If the tokens of one kind differ from each other in format, then this leads to an incorrect clustering of some tokens.

## 4. The proposed framework

In this section, we discuss and present the components of the proposed framework for extracting information from semi-structured web data sources. This framework consists of three components, namely: (i) Query Interface (QI), (ii) Information Extraction (IE), and (iii) Relevant Information Analyzer (RIA) as shown in Figure 2. In the following we discuss each of this component in details.

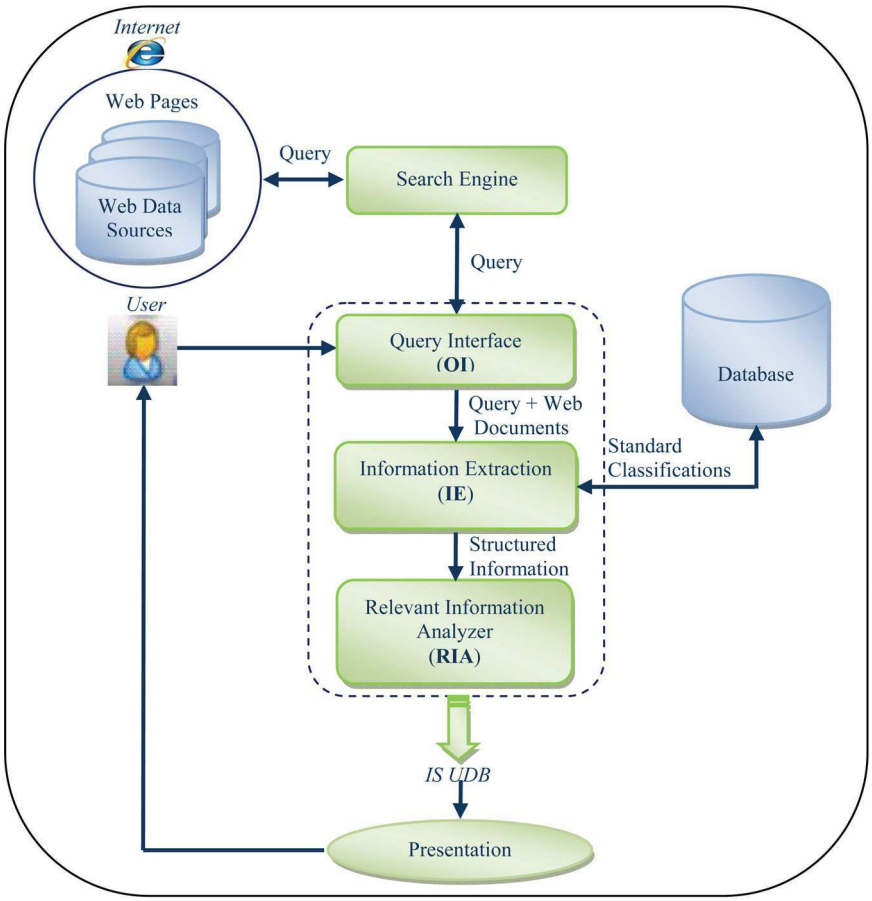


Fig. 2. The proposed framework

### 4.1 The Query Interface (QI)

The Query Interface is the key entry to the web and tool for accessing information. A user writes a product name (query) in the Query Interface, and the query is send to a search engine which searches the web data sources. The results of the query and web documents are saved in folders by the Query Interface as HTM files. Example of a simple query is shown in Figure 3.

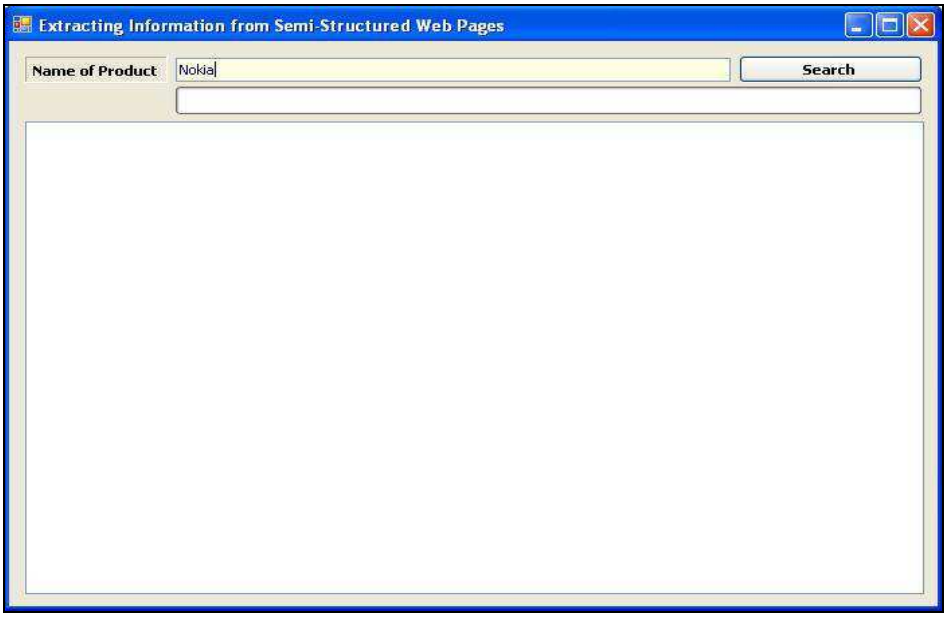


Fig. 3. The Query Interface

Figure 4 illustrates examples of the results of a query and the web documents which are stored in folders.

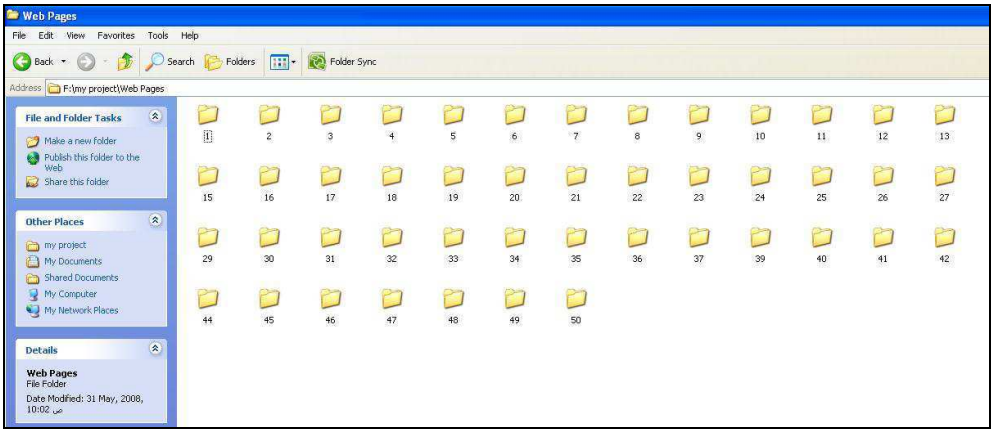


Fig. 4. Examples of results of a query and web documents

### 4.2 Information Extraction (IE)

The IE extracts and classifies the web pages that are stored in the folders, and converts them into text form. The details steps are discussed below.

Step 1: Based on the standard classification of Nokia products such as General, Size, Display, Ringtones, Memory, Data, Features, and Battery (Guntis Arnicans & Girts Karnitis 2006; Domenico Beneventano & Stefania Magnani 2004) (the attributes are shown in Figure 5) which is stored in database, IE extracts and classifies the web pages. Each kind of product is classified depending on the attributes. Figure 6 illustrates an example of the source code of a web page. IE extracts and classifies only the texts which are found between the tag "<TABLE>" and "</TABLE>". Figure 7 shows an example of information which is saved in an array by IE for matching this information with the standard classification of Nokia products. It illustrates the sub attributes and values of the sub attributes as shown in Figure 8 saved with the symbols "-" and ":" in an array. For example, the sub attribute Brand is saved with the symbol "-" which denotes a sub attribute (web page) and the value Nokia with the symbol ":" which denotes the value of a sub attribute (web page). IE ignores all texts which are not related to the standard classification, that are used for programming HTML web pages such as cellspacing, TBODY, TR, TD, row, href, >, <, /, etc.

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

Fig. 5. The standard classification of Nokia products

```
<DIV id=pricerunner>
<TABLE style="TEXT-ALIGN: left" cellSpacing=0 cellPadding=0>
 <TBODY>
 <TR>
  <TD class=spec_item>Brand </TD>
  <TD>Nokia </TD></TR>
 <TR>
  <TD class=spec_item>Type </TD>
  <TD>6212 classic </TD></TR>
 <TR>
  <TD class=spec_item>Form factor </TD>
  <TD>Candybar </TD></TR>
 <TR>
  <TD class=spec_item>Color </TD>
  <TD>Black </TD></TR>
 <TR>
```

Fig. 6. Example of the source code of a web page

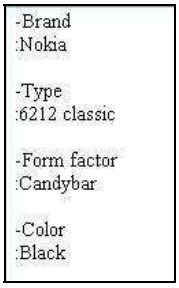


Fig. 7. The sub attributes (web page) and values of sub attributes (web page) shown in Figure 7 saved in an array

Step 2: Next IE converts the extracted and classified web page into text form. Figure 8 illustrates the example of the extracted sub attributes and values of the sub attributes, where each line begins with the index of an attribute (the standard classification) that is matched. For example, IE saves the sub attribute *weight* with the index of the attribute *Size*. 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 9 which illustrates the example of the extracted attributes and sub attributes that are shown in Figure 8 after grouping them based on the index number.
Figure 10 shows the web pages which are extracted, classified, and converted into text form by IE. The texts begin from text 7 until text 37, the IE ignores the web pages which are not related to Nokia products.

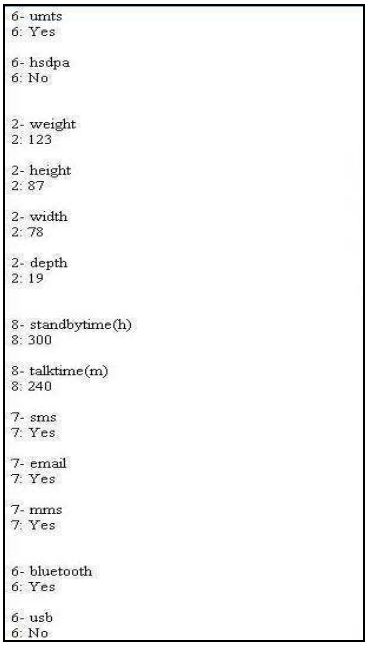


Fig. 8. The attributes of a web page in a text file with index number of an attribute (the standard classification)

2* SIZE
- weight
  : 123
- height
  : 87
- width
  : 78
- depth
  : 19

3* DISPLAY
- displaywidth
  : 128
- displayheight
  : 160
- lcdsize
  : NA
- seconddisplay
  : NA

4* RINGTONES
- voicedialing
  : Yes

5* MEMORY
- memory
  : NA

6* DATA
- gprs
  : Yes
- umts
  : Yes
- hsdpa
  : No
- bluetooth
  : Yes
- usb
  : No

7* FEATURES
- sms

Fig. 9. The attributes of a web page, sub attributes of a web page, and values of sub attributes in a text file after grouping

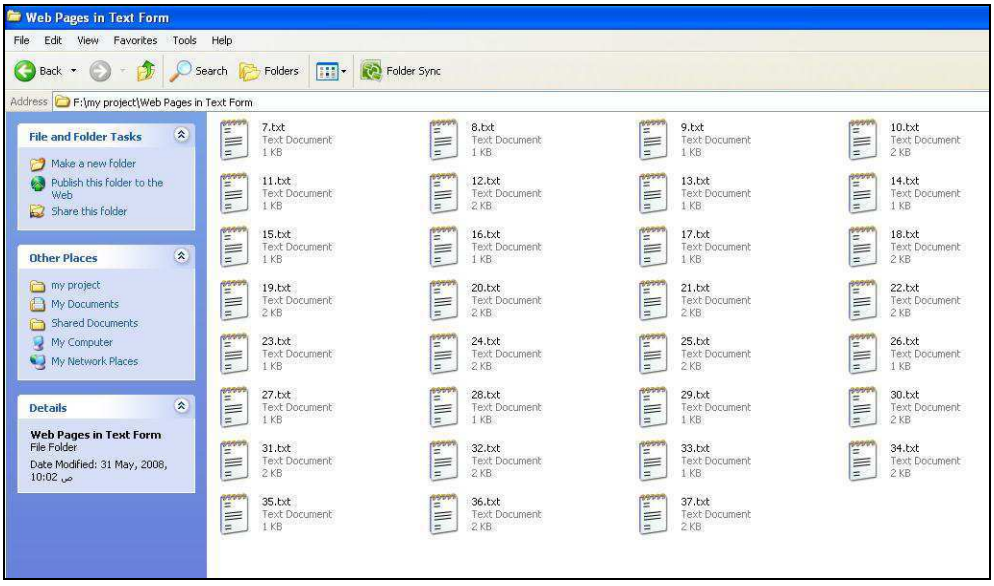


Fig. 10. Web pages in text form

Step 3: Next the IE converts the text (web pages in text form) to structured information. The IE counts the number of extracted sub attributes in each text form, and saves the results in a table. Figure 11 illustrates the number of extracted sub attributes in each text form.

| Name of text form | Number of extracted sub attributes in text form |
|---|---|
| Text 7 | 24 |
| Text 8 | 13 |
| Text 9 | 12 |
| Text 10 | 13 |
| Text 11 | 23 |
| Text 12 | 21 |
| Text 20 | 42 |
| Text 21 | 30 |
| Text 22 | 33 |
| Text 23 | 1 |
| Text 24 | 29 |
| Text 25 | 37 |
| Text 26 | 22 |
| Text 29 | 32 |
| … | … |
| Text 37 | 28 |

Fig. 11. The structured information

### 4.3 Relevant Information Analyzer (RIA)

The function of RIA is to determine the relevant information extracted from Information Extraction (IE) based on the number of attributes available in each text form. The steps performed by RIA are presented below.

Step 1: RIA receives the structured information from IE. Step 2: RIA determines the relevant information extracted from IE based on the number of sub attributes in each text. For example, text 23 has one sub attribute, RIA deletes text 23 because of small number of sub attributes. Sometimes one text (web pages in text form) has the same information found in other text or lesser for the same product, for example text 8 and text 10 have the same number of sub attributes. In this case, RIA deletes one of the texts for reducing the space of storage in the universal database (IS UDB) (Guntis Arnicans & Girts Karnitis 2006). IS UDB receives the rest of the texts from RIA and saved them in universal database. Figure 12 illustrates the results which are displayed to user. A user clicks on any products in the form (Figure 12) then the web sites which have the information about the product appears in a list box (Figure 12).

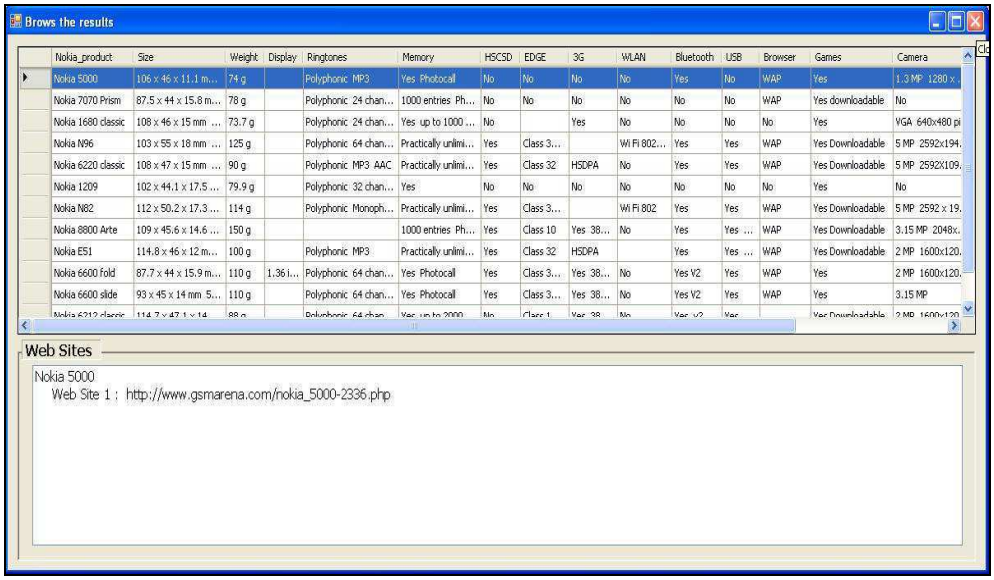


Fig. 12. Browsing the results to user

## 5. Conclusion

In this chapter, we proposed a framework to search and extract information from different web data sources. The proposed framework provides facilities to the user during search. A user does not need to visit the homepages of companies to get the information about any product, just write the name of product in the Query Interface (QI) and the framework searches all the available web pages related to the text which the user writes in the Query Interface (QI), and the user gets the information with little efforts.

## 6. References

B. Chandrasokaran, John R. Josophson, and V. Richard Bonjamins (1999). What are Ontologies, and Why Do We Need Them?, *Journal of IEEE Intelligent Systems and Their Applications*, Vol. 14, Issue. 1, pp. 20-26.

Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled F. Shaalan (2006). A Survey of Web Information Extraction Systems. *Journal of IEEE Transaction on Knowledge and Data Engineering*, Vol. 18, Issue 10, (Oct. 2006) pp. 1411-1428, ISSN: 1041-4347.

David Buttler, Ling Liu, and Calton Pu. 2001. A Fully Automated Object Extraction System for the World Wide Web, *Proceedings of the 21st International Conference on Distributed Computing Systems, Georgia Institute of Technology*, ICDCS, pp. 361-370, ISBN: 0-7695-1077-9, 2001, USA..

Domenico Beneventano and Stefania Magnani (2004). A Framework for the Classification and the Reclassification of Electronic Catalogs, *Proceedings of the 2004 ACM*

*Symposium on Applied Computing*, pp. 784-788, ISBN: 1-58113-812-1, Nicosia, 2004, Cyprus.

Fatima Ashraf, Tansel Ozyer, and Reda Alhajj (2008). Employing Clustering Techniques for Automatic Information Extraction from HTML Documents. *Journal of IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 38, (Sept. 2008) pp. 660-673, ISSN: 1094-6977.

Fatima Ashraf and Reda Alhajj (2007). ClusTex: Information Extraction from HTML Pages, *Proceedings of the 21st. International Conference on Advanced Information Networking and Applications Workshops (AINAW'O7)*, Vol. 1, pp. 355-360, ISBN: 978-0-7695-2847-2.

Guntis Arnicans and Girts Karnitis (2006). Intelligent Integration of Information from Semi-Structured Web Data Sources on the Base of Ontology and Meta-Models, *Proceedings of the 7th International Baltic Conference*, pp. 177-186, ISBN: 1-4244-0345-6, Vilnius, July 2006, Latvia University, Riga.

Horacio Saggion, Adam Funk, Diana Maynard, and Kalina Bontcheva (2008). Ontology-based Information Extraction for Business Intelligence, In: *Lecture Notes in Computer Science*, pp. 843-856, Springer Berlin, Heidelberg, ISSN: 0302-9743 (Print) 1611-3349 (Online).

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, *Proceedings of ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 1, pp. 368-371, ISBN: 978-0-7695-3496-1.

Katharina Kaiser and Silvia Miksch (2007). Modeling Treatment Processes using Information Extraction, In: *Advanced Computational Intelligence Paradigms in Healthcare – 1*, Vol. 48/2007, pp. 189-224, Springer Berlin, Heidelberg, ISSN: 1860-949X (Print) 1860-9503 (Online).

Kostyantyn Shchekotykhin, Dietmar Jannach, and Gerhard Friedrich (2007). Clustering Web Documents with Tables for Information Extraction, *Proceedings of the 4th International Conference on Knowledge Capture*, Canada, pp. 169-170, ISBN: 978-1-59593-643-1.

Line Eikvil (1999). Information Extraction from World Wide Web: A Survey, Norwegian Computing Center, ISBN: 82-539-0429-0.

Man I. Lam and Zhiguo Gong (2005). Web Information Extraction, *Proceedings of IEEE International Conference on Information Acquisition*, pp. 6, ISBN: 0-7803-9303-1, University of Macau, Macao, July 2005, China.

Nicholas Kushmerick (2003). Finite-State Approaches to Web Information Extraction, In: *Information Extraction in the Web Era*, ed. Maria Teresa Pazienza, pp. 77-91, Springer-Verlag Berlin Heidelberg, ISSN: 0302-9743.

Rik De Busser (2006). Information Extraction and Information Technology, In: *Information Extraction: Algorithms and Prospects in a Retrieval Context*, ed. Marie-Francine Moens, pp. 1-22, Springer Netherlands, ISBN: 978-1-4020-4987-3 (Print) 978-1-4020-4993-4 (Online).

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, Vol. 10, Issue 2, (March 05, 2007) pp. 157-179, ISSN: 1386-145X (Print) 1573-1413 (Online).

Stefan Clepce, Sebastian Schmidt, and Herbert Stoyan (2007). A Hybrid Approach to Product Information Extraction on the Web, In: *Advances in Intelligent Web Mastering*, pp. 68-73. Springer Berlin, Heidelberg, ISBN: 978-3-540-72574-9.

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, *Proceedings of ISIE, IEEE International Symposium*, Vol. 1, pp. 338-342, ISBN: 0-7803-7090-2, Pusin, June 2001, Korea.

Wolfgang Gatterbauer, Paul Bohunsky, Marcus Herzog, Bernhard Krupl, and Bernhard Pollak (2007). Towards Domain-independent Information Extraction from Web Tables, *Proceedings of the 16th International Conference on World Wide Web*, Canada, pp. 71-80, ISBN: 978-1-59593-654-7.


**Convergence and Hybrid Information Technologies**

Edited by Marius Crisan

ISBN 978-953-307-068-1
Hard cover, 426 pages
**Publisher** InTech
**Published online** 01, March, 2010
**Published in print edition** March, 2010

Starting a journey on the new path of converging information technologies is the aim of the present book. Extended on 27 chapters, the book provides the reader with some leading-edge research results regarding algorithms and information models, software frameworks, multimedia, information security, communication networks, and applications. Information technologies are only at the dawn of a massive transformation and adaptation to the complex demands of the new upcoming information society. It is not possible to achieve a thorough view of the field in one book. Nonetheless, the editor hopes that the book can at least offer the first step into the convergence domain of information technologies, and the reader will find it instructive and stimulating.



**InTech Europe**
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

**InTech China**
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821