PATTERN DISCOVERY AND ANALYSIS FROM SEMI-STRUCTURED DATA: THE AUTOMATIC REGENERATION OF PATTERNS

[1]ASMÂA ELOUERKHAOUI, [2]DRISS ABOUTAJDINE

[1]LRIT, Unité associée au CNRST n 29, Faculty of Sciences, Mohamed V-Agdal University,
B.P. 1014, Rabat 10000, Morocco.
Tel : (212) 6-69466359

[2]LRIT, Unité associée au CNRST n 29, Faculty of Sciences, Mohamed V-Agdal University,
B.P. 1014, Rabat 10000, Morocco.
Tel : (212) 3-7778973

E-mail: [1]elouerkhaoui@gmail.com, [2]aboutaj@fsr.ac.ma

ABSTRACT

Web-mining is the application of data mining techniques to extract knowledge from the world wide web, including web documents, hyper-links between documents, etc. Content-mining is the second step in Web data mining which mines web pages text to determine the relevance of the content to the search query. In 2008, we have proposed WIEBMat (Wrapper Induction Environment Based on Matrices) a new web content mining method[1]. This paper aims to optimize pattern discovery of WIEBMat. Effectively, the process of discovering frequent patterns from semi-structured data is one of the most important weaknesses in every web-mining technique .

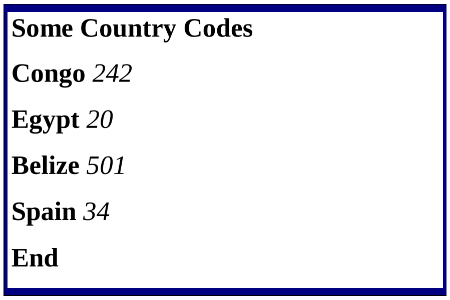
Keywords: *Web-Mining, Semi-Structured Data, Pattern Discovery.*

1. INTRODUCTION

WIEBMat is an application for extracting and gathering data from several web sources in order to generate a dedicated Meta-search engine. This approach is based on the analysis of resultspages structure that highly reduce and limit human intervention to input at maximum three instances examples in order to configure the meta-search engine with an existent web source.

A wrapper means both a proceeding for extracting tuples from a particular information source [3][4][5], and a function applied to query responses.

Taking as example the response to a specific query (fig.1) that allows displaying the countries names and their telephone codes. Initially, the web page contains structure errors, so as to be comprehensible by a learning machine expected to be converted into XML language. Add to this, the wrapper has to use delimiter strategy to indicate the relevant information context. Figure 2 below show it.



Some Country Codes

Congo *242*

Egypt *20*

Belize *501*

Spain *34*

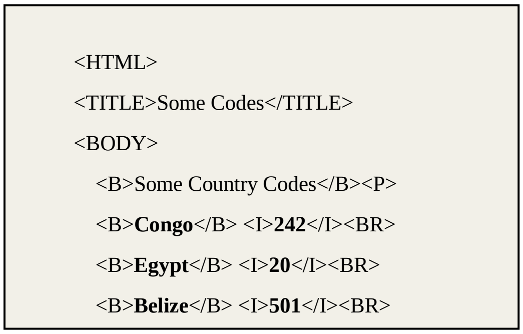
End

Figure 1: A Fictitious Internet Site Providing Information About Countries And Their Telephone Country Codes Given An Example Query Response.

From **figure 2**, it appears that this resource displays tuples by surrounding countries names with markups <B> and </B>, and country codes with markups <I> and </I>. The occurrences of this tuple are:

{(Congo, 242**); (**Egypt, 20**); (**Belize, 501**); (**Spain, 34**)}**

So, the selected wrapper relies on these four delimiters. But note that's just a simple definition of a wrapper induction system to extract information from web sources. Add to this, this simple left-right (LR) strategy, fails because not all occurrences of <B>... </B> indicate a country name. However, the string <P> can be used to distinguish the head of the page from the correct tuples.



```
<HTML>
<TITLE>Some Codes</TITLE>
<BODY>
  <B>Some Country Codes</B><P>
  <B>Congo</B> <I>242</I><BR>
  <B>Egypt</B> <I>20</I><BR>
  <B>Belize</B> <I>501</I><BR>
```

Figure 2: HTML Text From Which The Figure 1 Was Rendered

In short, an information extraction system consists in extracting a set of tuples from one or many query responses. Therefore, the information to be extracted is a set of fragments that have regular structure. Furthermore, this regularity (i.e., displaying countries' names in bold face and codes in italic) is not available as a machine-readable specification, but rather as a lack of normality aspect of the country/code resource. The term **'semi-structured'** is used simply to guide the intuitions about the kinds of information resource in which we are interested, and so we will not provide a precise definition.

Thereby, the process of a wrapper induction system is to parse web document represented as a query response to extract the relevant content while discarding the irrelevant ones.

2. WIEBMAT

WIEbMat (**W**rapper **I**nduction **E**nvironment based on **Ma**trices) [1][11][12][13][14][15] is an information extraction **(IE)** system. The main task of **IE** system is extracting structured information from unstructured/semi-structured documents. **WIEBMat** lies on an inductive environment as the **Gene/Clone** method [2][9] which we have proposed in earlier work.

Summarily, a web wrapper induction system can be either a system that need labeled examples or a system that can search for important information autonomously. In other words, the first kind of wrapper takes as input instance example of a chosen relation to extract the second technique, analyses the structure of a web document to output the repetitive data in a regular format so that the user can chose which one is relevant for him. **WIEBMat** as an inductive approach using instance examples on its input which are used to generate a pattern (This phase is certainly preceded by a pretreatment of web document). The pattern discovery is regular expressions describing context and delimiters of the relation to be extracted. The use of a matrix in WIEBMat approach allows the system to re-evaluate the calculated pattern in case where the matrix contains null elements.

Thus, we build an initial matrix which contains relevant and noisy data, by means of some calculated parameters we delete each row which contains at least one null element and we obtained at the end of the process a final matrix which contains only relevant data. We obtained very satisfying results by using this technique **(yield relevant data at a rate of almost 96%)**. The real problem which reduce this rate is the process of **XML-isation** (transforming a web pages coded in HTML language into XML format), that's why we have introduced a new parameter which calculates the capacity of the system to **XMLisate** a web page before performing information extraction (**XML-isation rate** must be higher than **0.37%**), thus, the experiments we conducted over several data sources have shown that web sources which has a rate of XML-isation lower than 0.37 cannot be efficient for WIEBMat algorithm). We implement this approach to create a Meta Search Engine called **BERG**. It uses in its back-office as many search engines as the user configure. **BERG** is a price comparison tool based on several commercial site (Cdiscount, Amazon, Pixmania, Priceminister, Rueducommerce).

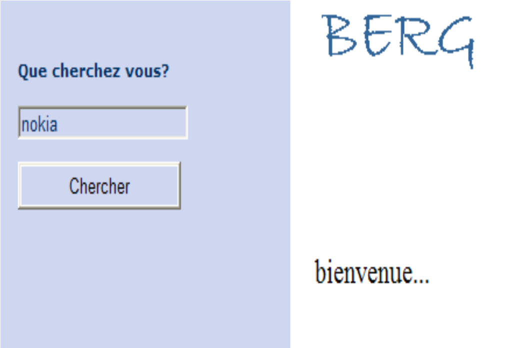
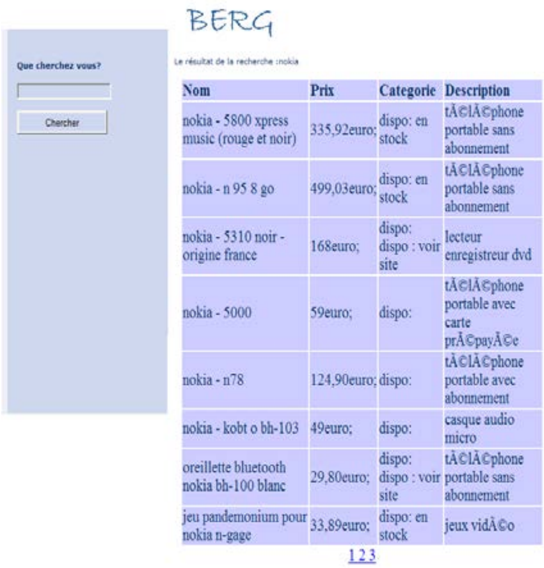


Figure 3: The Main Graphical Interface Of BERG



Le résultat de la recherche :nokia

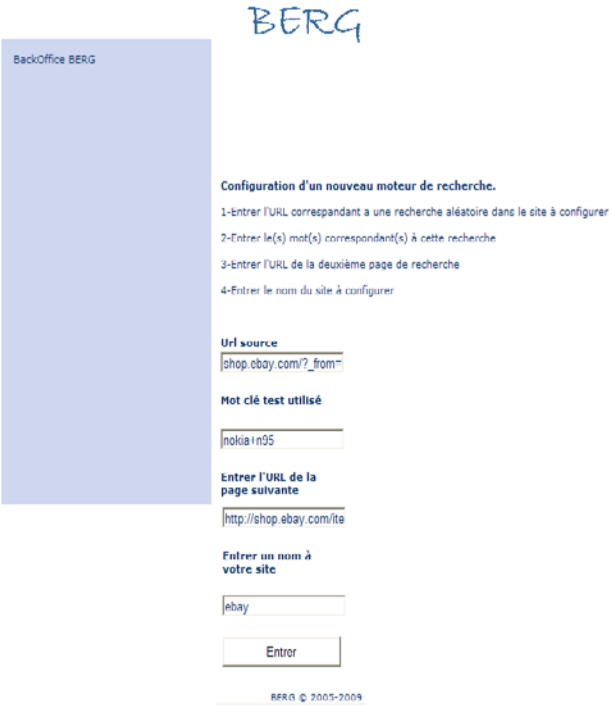
Nom	Prix	Categorie	Description
nokia - 5800 xpress music (rouge et noir)	335,92euro;	dispo: en stock	tÃ©lÃ©phone portable sans abonnement
nokia - n 95 8 go	499,03euro;	dispo: en stock	tÃ©lÃ©phone portable sans abonnement
nokia - 5310 noir - origine france	168euro;	dispo: dispo : voir site	lecteur enregistreur dvd
nokia - 5000	59euro;	dispo:	tÃ©lÃ©phone portable avec carte prÃ©payÃ©e
nokia - n78	124,90euro;	dispo:	tÃ©lÃ©phone portable avec abonnement
nokia - kobt o bh-103	49euro;	dispo:	casque audio micro
oreillette bluetooth nokia bh-100 blanc	29,80euro;	dispo: dispo : voir site	tÃ©lÃ©phone portable sans abonnement
jeu pandemonium pour nokia n-gage	33,89euro;	dispo: en stock	jeux vidÃ©o

1 2 3

Figure 4: A Response Web Page Of BERG

Once one or several keywords are entered in the text box shown in the **figure 1** we obtained as much as results found using the search engines configured in the back office. The utility of **BERG** lies in the conception of meta-search engines which fit to users need (eg. Filter results).

To configure a web source in the BERG's back-office we have to enter first of all the information shown in the third figure which are the source URL (Unified resource Locator) using at least two key words, and the URL of the second response web pages. Add to this we enter the name of the search engine to identify it and the number of results found in one single web page.



Configuration d'un nouveau moteur de recherche.

1-Entrer l'URL correspandant a une recherche aléatoire dans le site à configurer

2-Entrer le(s) mot(s) correspondant(s) à cette recherche

3-Entrer l'URL de la deuxième page de recherche

4-Entrer le nom du site à configurer

Url source: shop.ebay.com/?_from=

Mot clé test utilisé: nokia+n95

Entrer l'URL de la page suivante: http://shop.ebay.com/ite

Entrer un nom à votre site: ebay

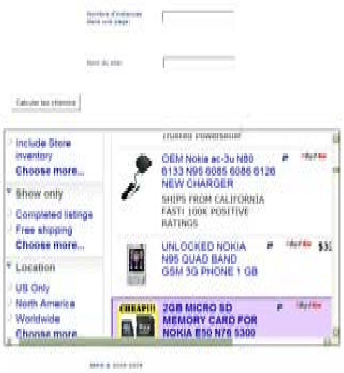
BERG © 2005-2009

Figure 5: First Graphical Interface Of BERG Back Office's

After that, we are redirected to the second graphical interface where we enter two example instances as shown in the figure four. This pattern may then be used to retrieve other instances.

The tuple used in this example is:

{(Name, price, Category, description)}



OEM Nokia ac-3u N80 6133 N95 6085 6086 6126 NEW CHARGER

SHIPS FROM CALIFORNIA FAST! 100K POSITIVE RATINGS

UNLOCKED NOKIA N95 QUAD BAND GSM 3G PHONE 1 GB

CHEAP!!! 2GB MICRO SD MEMORY CARD FOR NOKIA E50 N76 5300

Figure 6: Second Graphical Interface Of BERG Back Office's

As shown in this figure, we use a frame which contains a view of the first response web page obtained to take the pattern from it. The solution that we have proposed for the construction of an adapter for Web data sources leans on a set of example's instances of a relation to be extracted from these sources. The contexts of these instances are then extracted from a set of result pages. This method has several advantages with regard to the previous methods; it allows users to express simply their needs of information as a set relation instances to be extracted. The number of example instances to be given is often reduced: generally two or three instances for each source are enough.

3. PATTERN DISCOVERY AND ANALYSIS FROM SEMI-STRUCTURED DATA

There are several approaches which allows to obtain a wrapper, they are distinguished by the type of data they take as input. The most efficient ones are the inductive approaches; they consist on building a wrapper from a set of labeled examples pages introduced by the user, the structural approach which allows building patterns by analyzing the structure of per-formatting HTML pages to a comprehensible machine format as XML. Thus, this pattern is used to retrieve other instances

In this paper, we detect the patterns behavior so that we can improve them. This analysis has allow us to regenerating the pattern used to extract data instead of remain on the first calculated one. The idea is to use extracted occurrences from the generated pattern as new instances examples to reestablish it.

This technique is summarized in the following figure:

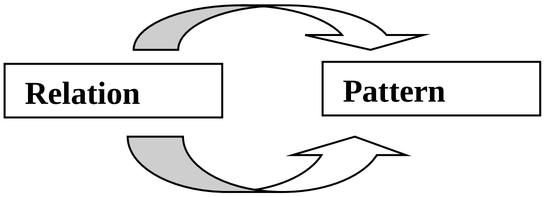


Figure 7: Pattern Analysis By Generating New Pattern From Extracted Occurrences (Duality Pattern-Relation)

The behavior of the pattern is associated with the evaluation of the F-measure (equation (1)). The F-measure combines the precision and recall rate.

$$F=\frac{2*Precision*Recall}{Precision+Recall}$$ Equation (1)

Precision (equation (2)) and Recall rates (equation (3)) are defined as follow:

$$Precision=\frac{tp}{tp+fp}$$ Equation (2)

$$Recall=\frac{tp}{tp+fn}$$ Equation (3)

Precision Recall measure the ratio between the quantity of extracted relevant data and the quantity of extracted one.

Recall measure the ratio between the quantity of extracted relevant data and the quantity of relevant one existing in the used document.

TP: Number of extracted relevant occurrences

FP: Number of extracted non-relevant occurrences

FN: Number of non-extracted relevant occurrences

Considering a d document used to extract data using a p Pattern, we applied **WIEBMat** approach pattern to extract the following relation: **{(**label**), (**price**), (**description**)}.**

Using initially two instance examples which are extracted from five preconfigured web sources 283 occurrences (the Precision rates, the Recall rates and the measure are evaluated below), we have applied the technique defined below. The found results are describde in the following diagram:

TABLE 1: EVALUATION OF INDICATOR : PRECISION/ RECALL/ F-MEASURE.

Precision	Recall	F-measure
0.41	0.98	0.578
0.51	0.97	0.668
0.53	0.96	0.682
0.62	0.95	0.75
0.68	0.94	0.789
0.73	0.93	0.89
0.87	0.92	0.817
0.91	0.91	0.91
0.92	0.89	0.9
0.93	0.86	0.89

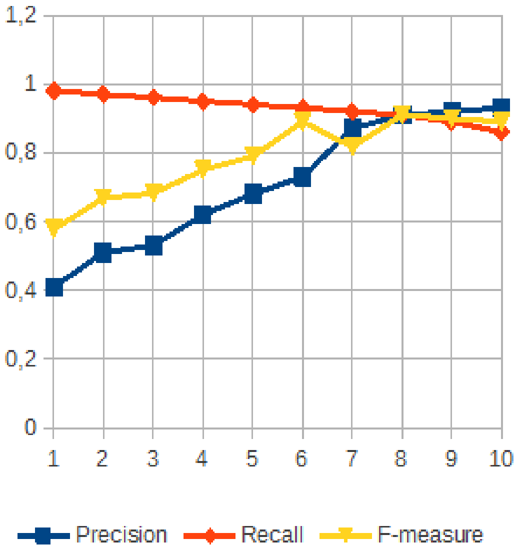


Figure 8: Evolution Of F-Measure According To Precision And Recall Rate

We concluded during these experiments that this servo mechanisms allow to calculate independently for each web source an optimal pattern retained when the system obtain the higher F-measure.

4. CONCLUSION

The automatic generation of wrappers that can extract relevant data objects embedded in semi-structured HTML pages expects that this mechanism generate a pattern which represents the structure of the data objects in the data-rich section. Thereafter, this pattern is used to extract data objects from a web page for a subsequent querying. In this paper, we evaluated the used pattern so that it can be steadily regenerated according to the highest F-measure found for the specific chosen relation and with the preconfigured web sources **(eg. F-measure=0.9, Precision=0.92 and Recall=0.89 for** www.ebay.com**).** This experience has demonstrated that an optimized pattern is not based on a high precision and recall rates, but on the duality between the relation to extract and the used pattern. Ultimately, the F-measure depends on the selected relation which designates the appropriate pattern.

REFRENCES:

[1] Asmâa Elouerkhaoui, Abdelaziz Sdigui Doukkali, “WIEBMat, a new information extraction system” IJCSNS International Journal of Computer System and Network Security, vol. 8, N.11, November 2008.

[2] El Habib BEN LAHMER, Abd Elaziz SDIGUI DOUKKALI, ElOuerkhaoui Asmaa. (2006). A new solution for data extraction: Gene/Clone method. IJCSNS volume 6.

[3] Kushmerick, N. , Weld, D. ,and Doorenbos, R. , Wrapper induction for information extraction. Proceeding of the 5th International conference on artificial intelligence (IJCAI), pp.729-735, 1992

[4] Kushmerick, N. , Weld, D. ,and Doorenbos, R. , Wrapper induction for information extraction. Proceeding of the 5th International conference on artificial intelligence (IJCAI), pp.729-735, 1992

[5] Conference paper or contributed volume Habegger, B., Extraction d’information à partir du Web, thèse pour obtenir le grade de docteur de l’Université de Nantes, 200

[6] El Habib BEN LAHMER, Abd Elaziz SDIGUI DOUKKALI, Mohamed OUMSIS. La Meta recherché générique: vers la génération des Meta moteurs de recherche. CopStic’03 Rabat, Mar

[7] El Habib BEN LAHMER, Abd Elaziz SDIGUI DOUKKALI, Mohamed OUMSIS. Towards an automatic extraction of data from half-structured documents. In ISCCSP2006.

[8] El Habib BEN LAHMER, Abd Elaziz SDIGUI DOUKKALI, Mohamed OUMSIS, 2004, WBERG, un Meta annuaire WAP, in isivc’04 Brest France.

[9] El Habib BEN LAHMER, Abd Elaziz SDIGUI DOUKKALI, ElOuerkhaoui Asmaa. (2006). A solution for data extraction by a new approach: The method of Gene/Clone. ICT4M: Kuala Lumpur, Malaysia.

[10] El Habib BEN LAHMER, Abd Elaziz SDIGUI DOUKKALI, ElOuerkhaoui Asmaa. (2006). BERG 2.2: a Meta search engine for on-line directories. ISCIT’06, Bangkok, Thailand.

[11] ElOuerkhaoui Asmaa, Abd Elaziz SDIGUI DOUKKALI. (2006). Comment render le contenu informationnel sur Internet intelligible. WOTIC’07, Rabat, Maroc.

[12] Elouerkhaoui Asmâa, Abdelaziz Sdigui Doukkali, 'Comment rendre le contenu informationnel sur Internet intelligible'. WOTIC'07, Rabat, Maroc.

[13] Elouerkhaoui Asmâa, Abdelaziz Sdigui Doukkali, 'Extraction d'information à partir d'Internet' AMINA'08, Monastir, Tunis.

[14] Elouerkhaoui Asmâa, Abdelaziz Sdigui Doukkali, 'Génération d'adaptateurs à partir de pages étiquetées'. WOTIC'09, Agadir, Maroc.

[15] Elouerkhaoui Asmâa, Abdelaziz Sdigui Doukkali, 'WIEBMat, a new information extraction system', JDTIC'09, Rabat, Maroc