An intelligent system for automatic data extraction in E-Commerce applications

Jesús Cardeñosa , Luis Iraola , Edmundo Tovar

Universidad Politécnica de Madrid, Facultad de Informática
Campus de Montegancedo. 28660 Boadilla del Monte. Madrid (SPAIN)
Tel: +34-91-3367436, Fax: +34-91-3524819, E-mail: carde@fi.upm.es

Abstract

One of the most frequent uses of Internet is data gathering. Data can be about many themes but perhaps one of the most demanded fields is the tourist information. Normally, databases that support these systems are maintained manually. However, there is other approach, that is, to extract data automatically, for instance, from textual public information existing in the Web.

This approach consists of extracting data from textual sources (public or not) and to serve them totally or partially to the user in the form that he/she wants. The obtained data can maintain automatically databases that support different systems as WAP mobile telephones, or commercial systems accessed by Natural Language Interfaces and others.

This process has three main actors. The first is the information itself that is present in a particular context. The second is the information supplier (extracting data from the existing information) and the third is the user or information searcher. This added value chain reuse and give value to existing data even in the case that these data were not tough for the last use by the use of the described technology.

The main advantage of this approach is that it makes independent the information source from the information user. This means that the original information belongs to a particular context, not necessarily the context of the user.

This paper will describe the application based on this approach developed by the authors in the FLEX ESPRIT IV n° EP29158 in the Work-package *"Knowledge Extraction & Data mining"* where the information captured from digital newspapers is extracted and reused in tourist information context..

Keywords:

Knowledge management; Knowledge reuse; Internet; Knowledge extraction

Introduction

It is well known the massive amount of contents available in Internet. This is the reason of almost all the current research efforts in Information Retrieval, the easy access to this information, the information filtering systems, the knowledge semantic structuration (ontologies) and many other technologies addressed to the intelligent information processing.

However, at this moment the applications are under the theoretical perspectives of many of these techniques. The integration of systems, technologies and components creates systems of enormous complexity with an average life under the capacity of change of Internet.

At the same time, new actors have appeared in this scenario, such as the "content providers". They have special difficulties to reach useful information for its users. The principal actors of this new phenomenon are: the information itself, the content provider and the user (or reader) of this information. (fig.1)

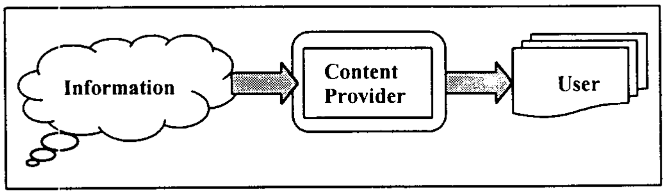

Figure 1 -Basic information flow

This context is complex because of the technologies involved and also because of the organization of the information flow. The content providers attempt to provide a variety of information, normally obtained from the information producers. This assumes that this information is accessible if the user knows the path to find it. And this is what makes difficult and less efficient the search of information.

The new systems for accessing information include complex natural language interfaces (with limited performance). The interfaces derived from the use of mobile telephones (WAP) using commands associated to the pieces of unknown information by the users are inefficient because it is very difficult to ask for information if the user does not known what is possible to ask to the system... Plato, mentioned more than two thousand years ago the problem "*A man cannot ask a question neither for what he already knows nor for what he does not. He does not need to ask a question about what he knows, and about what he does not know, he cannot ask it either, because he does not even know how to ask the question*". [1] Plato did not imagine the validity of his words...

In this complex context, new ideas are appearing. Perhaps one that includes all the others is the Knowledge Management [2], that is, the necessity to introduce new parameters in the basic information flow. These should permit at least to reuse knowledge from a domain into another [3]; to design interfaces really oriented to the user and not to the information [4] (the problem is not that they incorporate speech, natural language or text commands, but how this information is offered to the user) and all the issues derived from the storage of information to be used in an effective way. Furthermore, there are content providers in concrete domains that could maintain their databases automatically, simply introducing in their systems agents addressed to search this information in public domains, where the information has not to be provided by the generator of information because it is already in the WWW. This idea has been developed by the authors of this paper in the recently finished ESPRIT IV Project P29158 "FLEX" [5]. Technology has been developed and tested within this project.

The "FLEX" approach

In the context mentioned above, an application was developed to explore the viability of obtaining information from textual public access resources. In this sense, it deals to use no concrete and elaborated information but the data contained in the pieces of information. These data are not normally affected to Intellectual Property Rights but in any case it is not the subject of this paper.

The information is not captured in an open way, but concrete information oriented to a target domain already defined are searched. This chosen domain for the experience was the Festival of Edinburgh events, but it could be any other. The number of events is high (almost 10000) and the usefulness of this information system is indubitable. In the real case the organization of this event had already its database to support the information offered to the user.. In our case, this permitted us to match the truthfulness of the information captured from public access textual sources. As textual resources, we used articles published in the Digital newspaper "The Scotsman's" [6] that was collaborator of this project.

More than 1000 articles were used in this experiment. In these articles (art criticism, events description and many other contents, the information system automatically maintained searched the information. The searcher agent [7] was based on a series of grammars oriented to find the required information. Once this information is in the Database, is accessible by the way of the different interfaces depending on the target application. To access to the information in this Database for a WAP application (Wireless Application Protocol) is simple and it supposes to maintain automatically updated information in short periods. In the case of the FLEX approach the concept of Data Mining as generation of knowledge from Data using techniques of context analysis and semantic analysis was applied and these support the access interface to this information in Natural Language, module also developed during this experience.

The basic idea that supported this experience is represented in the Fig. 2. Once that the information has been captured from public access sources (or any other) it is available by the use of technologies that are in general mature enough. This could increase clearly the effectiveness of the searches in the web, where most of the times, the user access to texts searching an specific information (data). Otherwise, the idea to obtain information for an specific domain from information coming from other domain has been also object of study under different perspectives [8] (See Fig.2)

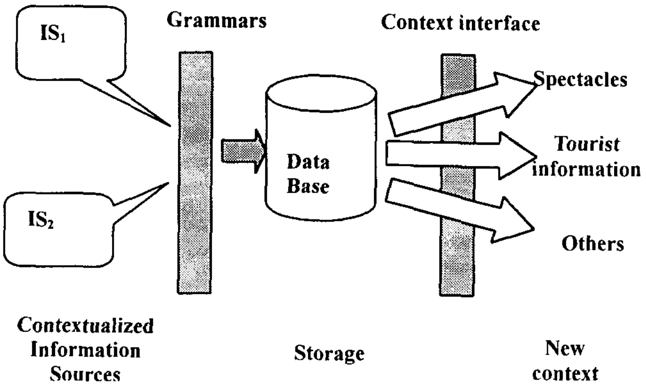

Fig.2 - Information management inter-domain flow

The FLEX application

In this paragraph, we will describe in detail the set of techniques and results obtained during the development of this application in the context of the FLEX project and the idea described before.

The textual information source

The work took as textual information source a corpus of 1146 newspaper articles of 1999, all of them related in one way or the other to the 1999 edition of the Edinburg International Festival (EIF). The collection varies from short notices of incoming events to large articles from the invited writers, with a majority of critical reviews of events.

These articles were revised and after this process, some parameters were determined for an easier allocation of the information. These parameters were:

1. ***Publication date***. Expressed in the format date/month/year (e.g. "14/8/1999"). Information present in all the articles.
2. ***Publication***. The name of the publisher, in our case the text "The Scotsman" appears in all the articles.
3. ***Headline***. The headline (or title) of the article.
4. ***Sub-headline***. Thematic classification of the article using just one word, like "Music" or "Theatre".
5. ***By-line***. The name of the author of the article.

6. ***Article.*** The body of the article.
7. ***Category***. Another classification, this one related with the area of the event reported, like "Preview fringe" or maybe the newspaper's section where the article belongs to, like "Review books".
8. ***Abstract.*** Summary of the article.
9. ***Rating***. A number, usually zero,. other values range from one to four.
10. ***Venue***. Information related with the place, time schedule and price of the event reviewed in the article.
11. ***Title***. The title of the event (the play, the concert) and/or the artist performing

Several sections were particularly relevant to our task, particularly the venue section where typically the place, time schedule and price were consigned. This is not to say that such information was always present. In fact most of the news articles failed to contain several information pieces quite relevant for our new domain (time, price and date most notoriously) but this can hardly be a surprise since the articles were not written for informing the readers about such details but mainly to provide a critical review of EIF's events.

The target domain

The amount and type of information needed to support the application of tourist data were defined according to the more typical information necessities of a visitor in this kind of events. Formats and "items" were defined even to be support of a mobile device connected to a WAP (Wireless Application Protocol) server. The following table summarizes the desired description of each event. (See table 1)

Technology

Our general approach to information extraction from collections of reasonably homogeneous items is to use grammars that take advantage of their regular linguistic characteristics for extracting the information pieces we are interested in. Furthermore, we advocate in favor of declarative, highly expressive grammatical formalisms in order to meet the overall requirements of cost-effectiveness and re-usability.

In our previous experience with some Scotsman's articles [4] we chose to write our grammars using the Definite Clause Grammar formalism, expanding their built-in expressiveness with meta-categorial operators that expressed optionality and repetition. For this experience, we have decided to expand a bit further the DCG formalism with operators that allow to express the occurrence of gaps or discontinuities in the definition of the categories. The newly expanded formalism has allowed us writing even more concise and declarative grammars.

Table 1 - Information structure in the target domain

Item	**Desired content**	**Format**	**Examples**
Title	Title of the event.	A string of characters. (CS)	"Shylock", "The princess and the tepee".
Genre	Artistic genre of the event, such as theatre, dance,	A genre code formed by four letters.	"thea" for theatre, "danc" for dance.
Festival	Within the EIF, different sections exist.	C.S.naming the section.	"Fringe", "Edinburgh Book Festival".
Person	Name of a person directly related with the even.	C.S. with the full name of the person.	"Agnes Martin", "Judy "Garland".
Venue	Name of the venue where the event takes place.	C.S. with the full name of the venue.	"King's Theatre", "Usher Hall".
Address	The address of the venue.	C.S. with the address.	"Charlotte Square", "5 Lothian Road".
Time	The starting time of the event.	A six-C.S. HHMMDD, with the hour, minutes, seconds	"153000" for half past three p.m. "190000" for seven o'clock p.m.
Price	The price o the seat, usually the average price.	An integer number or a floating point number. Two decimal numbers will be used, and the operating system decimal part separator will be used.	"12,99", "25".
Date	The date when the event happens.	A eight-character string YYYYMMDD, containing the year, month, and day.	"19990813" for August the 13th, 1999.

These grammars have been applied jointly with the following lexical resources:

- Oxford Advanced Learner's Dictionary of Current English.
- List of first personal proper names.
- Lists of English names of months and days.

Several auxiliary modules have been also used:

- Division into sentences.
- Tokenisation (implemented as a DCG).
- Lemmatisation.
- Procedures for overall control of the process.

These modules had been developed in Prolog, in tight integration with the DCG grammars. Most of these grammars are simple syntactic pattern recognizers that rely on orthographic and lexical features for detecting dates, time schedules, prices and addresses in the venue:

```
venue_info(Venue,Place,Date,Time,Price)-->
dcg_opt(date(Date))
dcg_opt(time(Time))
venue(Venue)
dcg_opt(location(Place))
dcg_opt(price(Price)).
date(Date)   --> dcg_gap(show_date(Date)).
time(Time)   --> dcg_gap(show_time(Time)).
venue(Venue) --> dcg_gap(venue_name(Venue)).
location(Place) --> dcg_gap(address(Place)).
price(Price)  --> dcg_gap(show_price(Price)).
```

This code fragment shows the use of the optionality operator (dcg_opt/1) when defining a legal venue section; only the name of the venue is required, the date, time, location and price are all optional. It also shows the use of the gap operator (dcg_gap/1), which takes from the beginning of the input list those tokens not recognized by the category and put them behind. The application of the gap operator to the five categories expected in the venue section we effectively make it order independent, since any token unrecognized by the first category (date) will be passed to the next one. In the following we will describe those operators that offered special difficulties.

Title and artist sections

After inspecting the corpus, we found that journalists tend to headline their articles with the title of the event the article is about. This is particularly the case when the event commented falls under the genres of theatre, comedy, opera, dance and film. Examples of such cases are in the table 2.

Table 2 – Examples of headlines with the title of the event

Article numbe	Section	Content
1	headline	Shylock
90	headline	Marlene! Live! Tonite!
164	headline	The Lunatic

On the other hand, articles about literary events like book presentations, lectures and so on contain headlines that do not name an event. In most cases, this is so for the sound reason that the article does not report on an event but on an interview, for instance. In other cases, the headline simply contain a phrase or sentence the journalist has considered adequate for the event he or she is reporting about. For example, article number 10 is entitled “What makes a country?” since it reports on several talks with different authors about nationalistic issues. And article number 13 has “Dark Danish commentator” as its title because its author found that phrase appealing for referring to Suzanne Brogger, a Danish novelist.

Musical events also cause problems since the articles reporting on such events may headline them with the names of the performers, as it happens in articles numbers 58 (headline: “Soile Isokoski and Marti Viitasalo”) and 56 (headline: “Vienna Philharmonic Orchestra”).

A pattern common enough to be recognized consists in putting in the headline the author or performer followed by the title of the show. Examples of this practice are in the table 3.

Table 3 – Examples of headlines composed of author+ title

Article number	Section	Content
851	headline	Geraldine McNulty: Greatest Hits
658	headline	Martin Bigpig: My Granny Was A Bearded Lady
487	headline	The Nualas -- The Big Shiny Dress Tour
700	headline	Egiku Hanayagi -- Japanese Dances

In order to recognize the pattern, we have written a simple grammar that detects the punctuation characters that may divide performers from title.

```
headline(NameL,Title) --> name_sequence(NameL),separator, title(Title).
headline([],Title) --> title(Title).
Separator --> dcg_around([blank], [punct(':')]).
Separator --> dcg_around([blank], ([symbol('-')], [symbol('-')])).
Separator --> dcg_around([blank], [symbol('-')]).
Separator --> dcg_around([blank], [symbol('/')]).
```

```
name_sequence([N|T]) -->
 proper_name(Name)
 {\+false_proper_name(Name) }
 dcg_star(X^(name_separator,proper_name(X)),T).
name_separator --> dcg_around([blank],[word(and,_)]).
name_separator --> dcg_around([blank],[punct(',')]).
```

Although this grammar produced many correct results, it also induced some errors. Occasionally, the writer has used the colon for other purposes ("O Caledonia: Sir Walter Scott and the Creation of Scotland", article 9), or put a remark after the title ("Labyrinth -- The Mystery of the Monster in the Maze"). In some other cases, it is really impossible to discern if the element separated is the name of the performers or just a remark ("World Drama: The Power of Seven", article 1118). All in all, we have found these mistakes (or possible mistakes) no so frequent and worth paying.

The separation between performers and show title has not been made in absence of those punctuation indicators. This results in titles that contain both elements, like for instance "The 4 Noels present The Magnificent Seventeen" (article 490), or "Andre Vincent is Cheeky" (article 650).

The "people involved" section

This information item was originally designed for storing the names of the people involved in an event, such as the performers or the authors of the event. However, it is not uncommon for theatre and music events to have as performers groups such as "Ensemble Modern" (article 464), "Genghis Blues" (140) and "Giralda Theatre Company" (512). Therefore, we have looked for both personal names and group names when filling up this item.

There is no explicit article's section where the people involved in an event is written down. This kind of information is contained within the article proper, interspersed with other information pieces the author has considered worth mentioning. Our strategy for extracting the names of relevant people or groups has consisted in mining the abstract and body sections, looking for "promising" proper names and then selecting those ones that appear at least two times in these sections. We assume that names occurring more frequently are those of the people or groups actually involved in the event. On the contrary, names that appear just one time are disregarded as non-relevant enough to be extracted. In order to implement our simple extraction strategy, we have had to cope with several difficulties:

- Our definition of a promising proper name has to be flexible enough for accepting names of groups as the ones aforementioned and also names of foreign people. From our previous experience on author extraction, we have collected a fairly comprehensive list of first personal proper names (around twelve thousand), but this list covers mostly English-language first names that we have found common in our corpus non-English first names. So, we have defined as promising proper name any sequence of capitalized words, and then we have to filter out unwanted candidates.
- Our relaxed definition of promising proper name accepts as proper names capitalized words that are not proper names. Beginning-of-sentence words, month and day names, national adjectives, etc have to be filtered out as a first step before selecting the most frequent names.
- When reviewing a play or a film, it is quite common to find in the article references to the characters appearing in the event. These character names can appear several times, confusing our strategy of picking up the most frequent names. We have avoided (most) character names by extracting only those names that are at least two words long.
- While relevant personal names are generally introduced the first time mentioning the first name and the family name (so, at least two words long), subsequent references are usually made using just the family name (so, just one word). Our counting procedure has been refined for coping with this phenomenon and at the same time avoiding character names.
- Although both the abstract and the body of the article had been mined for extracting as many promising proper names as possible, there are cases when no name is mentioned at least twice. In such cases, we have relied in a fall back strategy consisting in extracting all the names that are two words long and its first word is a known first personal name.
- Besides mining abstract and body, the performers detected in the headline (title) are also included as people related to the event.

Even after improving our strategy along these lines, its intrinsic simplicity makes it vulnerable to several mistakes:

1. Making possible the detection of group names also opens the door to other proper names, typically to venue names ("Abbey Theatre", article number 4) and play names ("The Speculator", article 3), but also to other names ("Stanford University", article 44, "King Arthur" article 47).
2. The requirement of at least two occurrences for being considered a relevant name discards many names that are relevant but got mentioned just once. This failure is common, much commoner than its opposite: to extract as relevant a name mentioned at least twice but anyway non-relevant one for the event. Though not so common, this type of mistake is of course present in our results: "Erwin Shrödinger" (article 422) or "Salvador Dalí" (articles 562, 725 and 1071).
3. Character names are mostly banned, but some of them do look like real-people names: "Sherlock Holmes" (articles 452 and 453).

Some of these known weaknesses can be mitigated adding more knowledge to the extraction process. For instance, venue places can be filtered out using a base of known venues. We have implemented a similar solution compiling a small list of words characteristic of venues and of the EIF as such, and then filtering out those names that contain one of such words. Other mistakes are much more difficult to avoid and the cost of devising an automatic procedure that cope with them may be as costly as performing a manual extraction. All things considered, our strategy is a compromise between simplicity and accuracy. Although it can be improved, the improvements are almost marginal in the whole corpus.

The results obtained from the 1146 articles looking for the information pieces described in this section can be summarized in the table 4. For each information item, the table shows the number of items extracted from the whole corpus:

Table 4 – Results of the process

Inform. Item	Items extracted	Remarks
Title	1000	Besides the 145 articles rejected because lack of enough information about the venue, 1 article was rejected because of empty title.
Genre	1000	All events extracted have been assigned a genre, "unknown" in cases.
Festival	1000	All events extracted have been assigned a festival.
Person	1646	
Venue	1000	145 articles of the initial corpus of 1146 articles were rejected because lack of enough information about the venue.
Date	166	
Time	14	
Price	19	
Address	102	

Technical details

In order to give a rough impression of the technical characteristics of the implemented techniques for dealing with the corpus, we give some figures:

- Number of software modules.- 8 Prolog code modules, classified in the following categories: 6 DGC Grammars (proper names, addresses, dates, headlines, venues and prices); 3 plain Prolog modules (generation of in-between days, selection of relevant personal proper names and overall control of the extraction process) and 4 knowledge bases (OALDCE, list of first personal names, genre codes, month and day names and canonical venues).
- CPU time consumed for the whole extraction process was of less than 10 minutes on a 400 Mhz Pentium II computer equipped with 128 MB of RAM
- Memory required during the extraction process of the whole corpus: 20Mb
- Software environment. Windows NT4. Prolog compiler SWI-Prolog 3.3

Conclusions

We have described a way to extract information from a domain and potentially in any format to be used for other purposes than the information originally produced. By reasons of space of the article, we have not included a complete description of all the items and grammars used on this application. [9]. In this case, the articles of general descriptions have been the real source of data to support an information site for these kind of applications. The final use of the data is independent of the original domain and the reuse of information is a fact that could mean an important reduction of time in the information maintenance and in general in the process of knowledge management for content providers. Once the algorithms have been defined and tested they can run continuously in the different collections of articles that have similar structure. Almost all the others produced were so. The experiment was done over a big enough collection to consider the experiment as completely valid. This structuration of the information was used in the project to be accessed by a natural language interface mainly developed by the company Sofware AG Spain, with very good results of accessibility by the user. We want to thanks its collaboration in the real application to the Scotsman Digital Newspaper.

References

[1] Platón. 1981. *Diálogos.* Selecciones Austral. Ed. Espasa Calpe.

[2] Cardeñosa J. 2001. *Intelligent systems in problem analysis in organizations.* Encyclopedia of Library in Information Sciences. Marcel & Dekker Ed.

[3] Neighbors J.M. 1997. The commercial applications of Domain Analysis. In Proceedings of the 8th Workshop on Institutionalizing Software Reuse.

[4] Cardeñosa J.; Iraola L.; Tovar E. 2001. Author extraction: A test Experience for flexible information extraction. pp 255-266. In Flexible Query Answering Systems . Henrik L. Larsen & alt. Ed. Advances in soft computing. Physica Verlag.

[5] Iraola L.; Cardeñosa J. 2000. Algorithms for data mining and for organizing knowledge. Deliverable 2.2. Technical Report. FLEX ESPRIT P29158. European Commission.

[6] The Scotsman. http://www.scotsman.uk

[7] Bradshaw J.M.; Suri N.; Cañas A.; Davis R.; Ford K.; Hoffman R.; Jeffers R.; Reichherzer T. 2001. Terraforming Cyberspace. *Computer*. Vol:pp 48-56. IEEE.

[8] Neumann, G.; Mazzini.., G. 1998. Domain-adaptive information extraction. Technical report, DFKI Saarbrucken.

[9] Iraola L.; Villa M. 2000. Report on the Edinburgh International Festival event descriptions extracted from Scotsman's articles. Deliverable 9.9.2. Technical Report. FLEX ESPRIT P29158. European Commission. 2000