Proceedings of the 5th WSEAS Int. Conf. on APPLIED INFORMATICS and COMMUNICATIONS, Malta, September 15-17, 2005 (pp215-220)

Public Data Retrieval with Software Agents for Business Intelligence

MIRJANA PEJIC BACH, NIKOLA VLAHOVIC, BLAZENKA KNEZEVIC
Graduate School for Economics and Business,
University of Zagreb,
Trg J.F.Kennedya 6, 10000 Zagreb,
CROATIA

Abstract: - The concept of business intelligence stress that firms should use all information available in order to increase efficiency of decision making. Usual technologies for business intelligence are data warehousing, OLAP and data mining, but they refer only to internal data.

Valuable external data that is freely available at the Web should also be used. In the paper we propose retrieval of public data with software agents for business intelligence. Software agent for retrieving data about stolen cars in Croatia is developed.

Data base of stolen cars in Croatia in last four years is created, and is used for better decision making in insurance company.

Keywords: - Business intelligence, Software agents, Data mining

1 Introduction

Business intelligence is a broad category that encloses technologies for gathering, storing, accessing and analyzing the data to increase the quality of decision making. The average analytics project implemented within businesses today is delivering a return on investment of 431% over five years. More than half of the implementations studied - 63% - delivered that payback in two years or less [10].

Although business intelligence usually refers to using data from data warehouse, and analyzing it with tools like on-line analytical processing and data mining, it also comprises usage of data from external sources. Lots of valuable data are available at the Web that could be used in order to increase efficiency of decision making. However, it is not easy to use data from the Web because they are usually in text form that is not suitable for analyzing. Also, data are often contained in .doc, .ppt or .pdf document types.

Software agents carry out tasks on behalf of another entity, and can be used for collecting data from the Web. The goal of the paper is to demonstrate that software agents are capable of collecting valuable data from the Web that could enhance information used for business intelligence systems. Software agent for collecting data about stolen cars from the Croatian Ministry of Internal Affairs web site is developed. Insurance company used the collected data for more realistic risk estimation.

2 Public data on Internet

2.1 Structure of web information

The Internet was originally developed as an infrastructure for research collaboration. The growing interest in publishing data using this media resulted in large amounts of data freely available to scientists but also business people and the general public. The introduction of World Wide Web has been primarily responsible for the explosive growth in Internet publishing and communication. It is estimated that during the period from 1998 to 2002 Internet content expanded by 218% [6].

Note that the Web content is twofold. A part of the information presented on the web is available as fixed web pages usually referred to as the surface web. A larger part of the information is contained in database driven websites thus creating dynamic web pages on demand. This information is usually referred to as the deep web or invisible web.

The size of surface web accessible to common search engines accounts for only 0,18% of the information contained in the hidden/deep web [7]. This makes finding relevant information a great challenge. Even though a number of agents for information filtering and web search engines exist most of the information contained in the deep web remains hidden.

2.2 World Wide Web and public records

Courts and government agencies at all levels of government - local, state, and federal - are increasingly making public records available on web sites. Some jurisdictions are just beginning, while others have done so since the mid-1990s [8].

There are two ways public records are accessible electronically. Some jurisdictions post them on their government web sites, thereby providing free or low-cost access to records. Government agencies and courts also sell their public files to commercial data compilers and information brokers. They in turn make them available on a fee basis, either via web sites or by special network hookups.

The reason that public records are published is that citizens can monitor government. Public records provide notice to all members of society of the official actions taken by government. They also provide notice of the "official" status of individuals and property. Making public records accessible to citizens via the Internet is a powerful way to arm people with the tools to keep government accountable.

Yet the public records also contain a great deal of information about individuals, often very sensitive information. These records can be gathered and transformed accordingly for the purpose of research or business decision making.

2.3 Public records available online in Croatia

Even though Croatian Web space commenced its expansion recently a number of public records are available. Some of the information is contained in the surface web and most of the information is contained in the invisible web. This is the list of some of the information available:

- National Bank currency exchange rates data & statistics – www.hnb.hr
- Zagreb Stock market reports- www.zse.hr
- Varazdin Stock market reports – www.vse.hr
- Phone directory – white & yellow pages – www.tportal.hr/imenik/
- Crime reports by the Ministry of internal affairs – www.mup.hr
- National official newspapers – www.nn.hr

3 Software agents

There have been numerous attempts to define the term agent in previous decades. Many of proposed definitions were rejected and subject to debate [9]. It was not until recently that a consensus definition was achieved. This general definition points out two key attributes agents have: autonomy of their actions and capability to act on behalf of someone or something. Having these two characteristics common to all agents in mind, a software agent can be defined as a computer program capable of independent (autonomous) action on behalf of its user or owner [11].

There are several dimensions to classify existing software agents.

Agents may be classified by their mobility, i.e. by their ability to move around a network. This yields the classes of *static* or *mobile* agents.

Agents can also be classed as either *deliberative* or *reactive*. Deliberative agents derive from the deliberative thinking paradigm which holds that agents possess an internal symbolic reasoning model, and they engage in planning and negotiation with other agents in order to achieve their goals. Reactive do not have any internal symbolic models of their environment, and they act using a stimulus/response type of behaviour by responding to the present state of the environment in which they are embedded [12].

Agents may also be classified along several attributes which ideally they should exhibit. Nwana & Ndumu have identified a minimal list of three such attributes [12]: autonomy, learning and cooperation. According to the extent of presence of these characteristics there are seven categories of agents [3]: collaborative agents, interface agents, mobile agents, information/Internet agents, reactive agents, hybrid agents and smart agents.

A specific group of software agents are agents used for information management. These information agents are mostly used to reduce the information overload in their working environments. Currently, most of information agents reside on servers and are used for Internet information tasks such as web search, information filtering, information retrieval, notification and information services:

1. Web search agents try to locate *web pages* specified by the user with the optimal recall (number of total relevant documents) with the highest possible precision.

2. Information filtering agents try to locate *content*, personalize it and deliver it to the user.
3. Information retrieval agents deliver this personalized information in a locally viewable format.
4. Notification agents inform their users if there is a change in the state of information of interest to the user such as content change of a particular Web page.
5. Information agents may also provide announcement services, direct mail services, financial services, job services, shopping services and so forth.

For the purpose of this paper a smart information retrieval agent was developed. This agent can deliver delegated information form a specific deep web resource published on the Internet in a form of a database, while ensuring consistency of the data as well as offering possible solutions for correcting observed inconsistencies. These capabilities will be discussed in the next chapter.

4 Web information agent for data acquisition

4.1 The process of collecting data with web information agent

Web information agent for data acquisition that is capable to retrieve information and organize it as a local database conduct 4 steps: accessing data, retrieving data, checking data record consistency, and preparing and storing data locally.

First step is to determine the access to the data. Usually data is stored according to the time it is generated using some sort of an indexing system. Indexing system is used to navigate through archived data. Agent should be able to recognize two types of data sources: real-time data source and historical data source. There are several ways of presenting dynamic data contained either on the surface web or the hidden web using different indexing systems (Fig. 1).

Depending on the type of the web page, agent locates pieces of information required for database storage (second step). Depending on the type of information agent can check data for errors and propose possible corrections (third step). After data is screened by the agent and any inconsistencies within data are corrected, agent constitutes a local database and stores information (fourth step). These steps are repeated iteratively until all data is retrieved and stored. Additionally agent can monitor changes in data and update these changes on a daily bases (i.e. append newly published data).

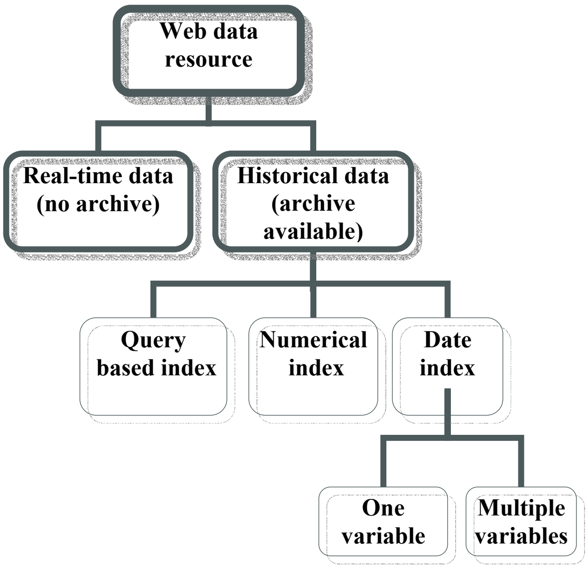


Figure 1. Indexing systems of the web information source

4.2 Collecting data about stolen cars

Data on cars thieves are published daily on the Croatian Ministry of Internal Affairs web site http://www.mup.hr. In addition, there is an archive of thefts during the last four years. Data on stolen cars are very useful for insurance companies because they help them assess which brands and regions are more risky, so that they can charge higher premiums to their owners. For every date, an archive page is created and added to the collection of web pages. The name of the page is used as a date index distinguishing each page in the collection.

A date index can consist of one or multiple variables (as shown in Fig. 1). Multiple variables date index usually uses separate variables for year, month and day. Example is daily.asp?dan=9&mjesec=6&godina=2004. One variable data index can have a number of forms. Most used ones are: DD-MM-(YY)YY, MM-DD-(YY)YY, (YY)YY-MM-DD. An example of a valid date index is 040621.jpg where digits are associated as YYMMDD.jpg. Investigation showed that there are valid forms (mentioned above) and invalid forms (such as D-M-(YY)YY). Agent should be capable to recognize invalid forms since in this case the archived data will not be complete. In this case agent should avoid errors in collected data by monitoring the data provider as if it published real time data.

In the case of car thefts, an invalid index is used. An example is 2004122.html where digits are not strictly associated with dates. The index in this example can stand for the date December 2nd, 2004 and for January 22nd, 2004. This means that data for January will no longer be fully available after December 1st (as it will be replaced by the data for December 2nd). This is the reason why database about car thieves collected by software agent cannot be complete. However, we think that it is better to have some data even with missing values instead of no data.

Information on a web page can be structured, semi-structured [1] or unstructured [2]. Unstructured content of a web page is represented by plain text. Gathering information from the unstructured text requires natural language processing methods. Semi-structured content of the web page consists of a well structured plain text. In the case of car thefts data is published as sentence that has always the same semi-structured form. One example is:

***Zagreb** – on February, 20th between 10,30 p.m. and 7 a.m. at the New Zagreb area from the address Katićev prilaz the car vehicle VW golf was stolen with license plates ZG 8617-AT, owned by B.B., estimated for 100 thousand kuna.*

These short reports have a structured scheme given bellow:

***(Police station region)** – (date) (time) (location) (item in question – brand and type) (license plate number) (owner data) (value of the item).*

Each section of the report is separated by the stop-words. A list of typical stop-words (and stop-symbols) denoting beginnings and endings of each section is given to the agent during training sessions. Agent locates each section by the given stop-words and analyses each section. Based on the analysis agent can detect erroneous information or derive new information. For example, using information about the date agent derives the information about the day in the week the car was last seen.

Agent was trained to collect the data from the semi-structured reports as described above, and local database was constituted. The following data is collected:

- year
- region
- date when the car was last saw by the owner
- date when it was registered that the car was stolen
- day of the week when it was registered that the car was stolen
- time of the day when it was registered that the car was stolen
- exact location
- car brand
- car type
- license plates number
- owner initials
- value of the car

We continue to use the agent for monitor changes in data and update these changes on a daily bases.

4.3 Analyzing the data

Data about car thefts collected are used by one insurance company. Lots of different analysis was conducted, and even very simple analysis was useful. Risk indexes and association rules are applied, and it is planned to create data cube with Cognos in order to conduct OLAP.

Different risk indexes are calculated. For police affairs additional indexes, like day of the week index, and time of the day index. Insurance company found that brand risk index, type risk index, and region risk index were useful. Table 1 contains brand risk indexes. We can see that AAA is the most risky brand, which is expected because of its popularity and high value. Insurance company can charge higher brands for Audi owners than for owners of other brands that have the same value.

Table 1. Brand risk index

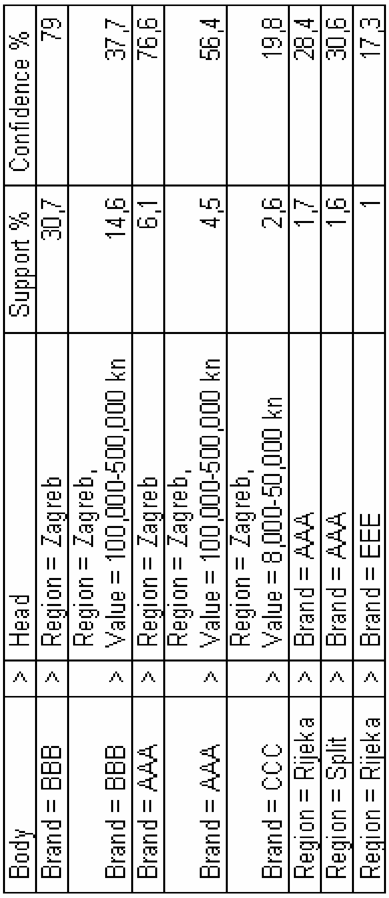
Brand	# of stolen / 1000 cars
AAA	6,7
BBB	5,0
CCC	5,0
DDD	4,5
EEE	2,0
FFF	1,2
GGG	0,9
KKK	0,8
MMM	0,7
PPP	0,6
QQQ	0,5
RRR	0,4

Data mining with association rules was conduced with the usage of software Statistica Data Miner in order to find relationships between brand, region and value. Association rules are undirected data mining method which goal is to discover associations between specific values of categorical variables in large data sets. An example of an association rule is: "30% of transactions that contain beer also contain diapers; 2% of all transactions contain both of these items".

Several association rules with three variables (brand, region, value) important for car thieves were detected (Table 2). For example, it was discovered that 30,7% of all thefts refers to brand BBB in Zagreb, 14,6% refers to brand BBB in Zagreb with value between 100,000 and 500,000 kn.

Rules discovered have rather low support. However, the low support top rules (e.g. some unusual combinations of some factors that have always caused some disease) may be very interesting if they have high confidence value [13]. For example, only 1,7% of all thefts refers to region Rijeka and brand AAA. However, this rule has high confidence of 28,4%, which is in fact conditional probability - that an observation that contains region Rijeka also contains brand AAA.

Table 2. Summary of association rules



Body	>	Head	Support %	Confidence %
Brand = BBB	>	Region = Zagreb	30,7	79

Brand = BBB	>	Region = Zagreb,
		Value = 100,000-500,000 kn
			14,6	37,7

Brand = AAA	>	Region = Zagreb	6,1	76,6

Brand = AAA	>	Region = Zagreb,
		Value = 100,000-500,000 kn
			4,5	56,4

Brand = CCC	>	Region = Zagreb,
		Value = 8,000-50,000 kn
			2,6	19,8

Region = Rijeka	>	Brand = AAA	1,7	28,4
Region = Split	>	Brand = AAA	1,6	30,6
Region = Rijeka	>	Brand = EEE	1	17,3

5 Conclusion

Business intelligence is a broad term that comprises all the technologies connected with data manipulation for better decision making. It usually refers to the data warehousing, OLAP and data mining. Usage of data from the Web retrieved by software agent for business intelligence was explored.

Goal of the paper was to demonstrate that software agents are capable for collecting valuable data from the Web. We have developed software agent for collecting data about stolen cars from the Croatian Ministry of Internal Affairs web site. The process of developing such agent is presented together with the analysis of the data collected. Insurance companies could benefit from the data collected in order to estimate risk more realistically.

References:

[1] Mohammadian M. intelligent Agents for Data Mining and Information Retrieval. Hershey: Idea Group Publishing, 2004.

[2] Liautaud B, Hammond M. e-Business Intelligence: turning information into knowledge into profit. New York: McFraw-Hill, 2001.

[3] Bradshaw JM. Software Agents. Menlo Park: AAAI, 1997.

4. Bigus JP, Bigus J. Constructing Intelligent Agents Using Java. New York: Wiley & Sons, 2001.

[5] Caglayan A, Harrison C. Agent Sourcebook. New York: Wiley & Sons, 1997.

[6] Neil E. Web Characterization. OCLC - Online Computer Library Center, 2004. http://www.oclc.org/research/projects/archive/wcp/stats/size.htm [31/01/2005]

[7] Lyman P, Varian HR. How much information? 2003. Regents of the University of California, 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm [31/01/2005]

[8] Givens B. Public Data on the Internet: Privacy Dilemma. Privacy Rights Clearinghouse/UCAN, 2002-2005. http://www.privacyrights.org/ar/onlinepubrecs.htm [28/01/2005]

[9] Foner LN. What's an Agent Anyway? A sociological case study. Agents Group, MIT Media Lab, 1993. http://foner.www.media.mit.edu/people/foner/Julia/Julia.html [27/09/2004]

[10] Vesset, D. Worldwide Business Intelligence Tools 2004-2008 Forecast. Framingham: IDC, 2004.

[11] Wooldridge M. An Introduction to Multy-Agent Systems. London: Wiley & Sons, 2002.

[12] Nwana H, Ndumu D. An Introduction to Agent Technology. BT Technology Journal 1996; 14(4): 55-67.

[13] Zhang, X, Li, J., Dong, G. Ramamohanarao, K., Qun, S. Efficient Mining of High Confidence Association Rules without Support Thresholds. Proceedings of PKDD 99 -- 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, Prague, September 1999.