State-of-the-art web data extraction systems for online business intelligence

Tomas Grigalis

Vilnius Gediminas Technical University, Faculty of Fundamental Sciences,
Department of Information Systems, Doctoral Student
Vilniaus Gedimino techniko universiteto, Fundamentinių mokslų fakulteto,
Informacinių sistemų katedros doktorantas
Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
E-mail: tomas.grigalis@vgtu.lt

Antanas Čenys

Vilnius Gediminas Technical University, Faculty of Fundamental Sciences,
Department of Information Systems, Prof. Habil. Doctor
Vilniaus Gedimino techniko universiteto, Fundamentinių mokslų fakulteto,
Informacinių sistemų katedros Prof. Habil. Daktaras
Saulėtekio al. 11, LT-10223 Vilnius, Lithuania
Tel. (+370 5) 274 5005
E-mail: antanas.cenys@vgtu.lt

The success of a company hinges on identifying and responding to competitive pressures. The main objective of online business intelligence is to collect valuable information from many Web sources to support decision making and thus gain competitive advantage. However, the online business intelligence presents non-trivial challenges to Web data extraction systems that must deal with technologically sophisticated modern Web pages where traditional manual programming approaches often fail. In this paper, we review commercially available state-of-the-art Web data extraction systems and their technological advances in the context of online business intelligence.

Keywords: *online business intelligence, Web data extraction, Web scraping*

1. Introduction

The Web is full of data. Never before have we witnessed such constantly increasing and at first glance easily accessible repository of data about everything. The size of the indexable Web, i.e. the Web sites which are considered for indexing by the major search engines, is thought to be at least 11.5 billion pages as of the end of January 2005 (Pis et al., 2005). Moreover, a 400 to 550 times larger amount of information resides in the *Deep Web* (Bergman, 2001). The term *Deep Web* refers to the content hidden behind the Web forms. In order to retrieve such content, a user has to interact with Web page forms and perform a meaningful submission. The prime examples of Deep Web are car advertising listings, real estate listings, statistical department databases, etc. Accessing information published on Web (or Deep Web) sites has been a long-standing challenge for the data

extraction community (Bohannon et al., 2012; Cafarella et al., 2011; Elmeleegy et al., 2011; Furche et al., 2012; Madhavan & Halevy, 2009).

Today's competitiveness dictates the need for business organizations to constantly seek timely and accurate information to support decision making and action. The information seeking process is usually defined as *business intelligence* (Fleisher & Bensoussan, 2003), and the Web is a nearly perfect source for it. More specifically, the term *business intelligence* can be referred to as (Lönnqvist & Pirttimäki, 2006):

1. Relevant information and knowledge describing the business environment, the organization itself, and its situation to its markets, customers, competitors, and economic issues.
2. An organized and systematic process by which organizations acquire, analyze, and disseminate information from both internal and external information sources significant for their business activities and for decision making.

Some other related terms include competitive intelligence, market intelligence, customer intelligence, competitor intelligence, strategic intelligence, and technical intelligence (Lönnqvist & Pirttimäki, 2006). Business intelligence provides valuable information to companies in a timely and easily consumed way and enhances the ability to reason and understand the meaning of it through, for example, discovery, analysis, and *ad hoc* querying (Lönnqvist & Pirttimäki, 2006). When dealing with data coming from online sources, i.e. the Web, the term *online business intelligence* is used.

There is an enormous amount of data on the Web, researchers have been studying the data extraction field for decades, businesses are striving for valuable data and are more than ready to pay for it; however, there is still no universal and easily deployable solution to fully leverage the data on the Web. Business intelligence systems seeking to collect valuable data from the Web must overcome many great challenges posed by the Web itself, such as finding appropriate data sources, determining their credibility, navigating technologically sophisticated Web sites, submitting meaningful Web form queries to retrieve the Deep Web content, extracting, cleansing, understanding and integrating data, and constantly dealing with heterogeneity at each step (Baumgartner et al., 2009; Cafarella et al., 2011; Cafarella et al., 2008; Madhavan et al., 2007).

The aim of this paper is to review commercially available state-or-the-art Web data extraction systems in the context of online business intelligence. First of all, to remove any ambiguity from the term *online business intelligence* and to be more specific, in Section 2 we list concrete business scenarios and applications where online business intelligence is, or can be, practically employed. Then, in Section 3, we describe the current challenges which arise when collecting and processing data from technologically sophisticated modern Web pages. In Section 4, we review the commercially available state-of-the-art Web data extraction systems that can be used for online business intelligence, and further we list some even more promising technological advances.

2. Real world online business intelligence scenarios

Given that most information on pricing, product availability, store locations, etc. is available on the Web, *online market intelligence* is becoming the most important form of business intelligence (Baumgartner et al.,

2009). Market intelligence is the ability to understand, analyze and assess the environment of a firm with customers, competitors and markets, and industries, that conduces strategic planning and help decision making (Juntarung & Ussahawanitchakit, 2008). Currently, almost every large retail company has online market intelligence needs for marketing and pricing (Baumgartner et al., 2009). Here, we list typical real world examples when businesses employ online business intelligence to monitor their market environment.

2.1. Background Check

Background screening is a profession that absolutely demands precision and timeliness. Erroneous background checks could result in stiff penalties as well as loss of business (Connotate, 2012). By background checking, companies are verifying the background of a customer or a business partner. For example, the background checking process may involve accessing courts' Web sites in hundreds of jurisdictions: an automated data extracting program must query each Deep Web court database to check if a particular person or a company has any ongoing legal battles or does have any active legal restrictions, etc. In addition to online court records, many professional associations post information on the Web. Online business intelligence can allow companies to aggregate data from multiple sources to perform comparisons and to verify the validity of credentials (Connotate, 2012), such as certificates, honors, etc.

2.2. Competitive & Pricing Intelligence

A company's success hinges on identifying and responding to today's hypercompetitive environment, especially in online settings. A company's challenge is to gather accurate competitive intelligence, analyze it and act as quickly as possible (Connotate, 2012). Consider, for example, three scenarios presented by Baumgartner et al. (2009):

- "an electronics retailer would like to get a comprehensive overview of the market in the form of a dashboard displaying daily information on price developments including shipping costs, pricing trends, and product mix changes by segment, product, geographical region, or competitor."
- "a supermarket chain wishes to be continually informed about their competitors' product prices. Moreover, they want to be immediately informed in case a competing supermarket chain issues a special offer or promotion. They need to react very quickly to price changes or new special discounts in order to maintain their competitive position. They also want to be informed as soon as new products show up on the market."
- "an online travel agency offering a best price guarantee needs to know at which prices the packages they offer are sold over the Web by competing travel agencies. Moreover, they wish to be informed about the average market price of each travel product they feature."

2.3. Compliance & Risk Management

Companies may want to be informed of updates on sanctions lists and regulations at the international, federal, and state levels. Automated online business intelligence makes it easier to ensure compliance with laws regulating rogue nations or organizations financing, to ascertain the legal integrity of a potential business partner, and to reduce exposure to financial fraud and identity theft

(Connotate, 2012). For example, a company working in car or car-parts trading business may want to check automatically if a particular vehicle or its part is not included in stolen property registers around the globe. Cost-effective management of compliance and risk is a complicated challenge, because precise source monitoring and Deep Web querying must be executed on-the-fly: all the automatically accessed data should be immediately extracted, cleaned, integrated and presented.

2.4. Customer Sentiment Analysis

Today, many customers are buying online and publicly sharing their user experience, opinions and buying preferences. In most cases, users express their opinion as comments, forum or social media post, tweets, etc. Analyzing customer sentiment is fundamental for maintaining a competitive edge in the delivery of goods and services (Connotate, 2012). Online business intelligence solutions should be able to access Facebook, Twitter, or any other social media Web site, automatically identify posts about a particular product, extract the text, execute a natural language processing and understand the sentiment. The same data extraction process can also be applied to hundreds of other sources, such as blogs, online forums, product review sites, YouTube, etc.

2.5. News & Content Aggregation

The media monitoring companies aggregate news articles and comments from many online Web sites. It is not a trivial problem to monitor hundreds of online sources which are usually heterogeneous in style, news format, and navigation. Furthermore, leading media monitoring companies should be able to automatically classify the collected news articles by their topic and to group articles describing the same event. The brand name mentioning monitoring is also an important task which should be executed on thousands of news articles. A proper online business intelligence solution should offer a scalable, automated technology to crawl and extract data from hundreds of thousands of news sites, archives, and corporate websites.

2.6. Financial Data Aggregation

According to Connotate (2012), "every moment of every day, political events, financial filings, corporate actions and many other market-moving events are posted on the Web. Detecting and communicating these events to the financial community in near real-time is essential to building and maintaining market share in the world of financial data. Speed and accuracy are paramount". New or updated financial data appear every day in a big variety of online sources, such as government data portals, news articles, companies' news feeds, even in social media. For example, hedge funds involved in algorithmic trading monitor thousands of sources in real time to immediately detect breaking news and swiftly sell or buy particular stocks. Even such trivial data as weather temperature are monitored across the globe, and in case of an unexpected drop in temperature the stocks of oil, gasoline, electricity or heater manufacturing companies can be bought in a matter of seconds. With the increasing wealth of information and content available on the Web, the opportunity to use it for a timely notification, analysis and enhanced decision making is unprecedented (Connotate, 2012).

3. Data extraction challenges posed by modern Web pages

The Web is evolving. To enhance the browsing experience, the modern Web pages rely

on a number of sophisticated technologies, such as Cascading Style Sheets (CSS), to separate the presentation style of Web pages from their content, Asynchronous JavaScript Requests (AJAX) to dynamically load the Web page content, client side scripting to modify the appearance of a Web page solely on client side, etc. Modern Web browsers are a complicated software that requires a considerable amount of computational power to visually render Web pages. Looking from the user perspective, all these technological advances of the Web enhance the browsing experience. On the other hand, they hinder Web data extraction.

To extract data today is not enough to simply download a Web page from a server and to analyze its source code. Modern Web pages are sometimes loaded incrementally using AJAX requests. JavaScript can modify the asynchronously received data according to some algorithms and only then display them on a page. For example, some modern Web sites use the AJAX technique for pagination: when the user clicks the "next page" link, a browser does not reload the whole page; instead, only required data are asynchronously retrieved from the Web server and displayed as the next page. Moreover, each step of navigation inside a Web site can be controlled by cookies and referrers. All these and many other features of a modern Web site for data extraction mean only one thing: each data extracting tool should fully emulate a modern Web browser. It is an incredibly hard task.

The Deep Web pages pose another challenge. In order to retrieve the information behind the Web forms, a data extracting system must execute meaningful queries. This means that it is becoming very difficult to algorithmically code the data extraction process, and the use of artificial intelligence techniques to understand and submit the Web forms is sometimes inevitable.

Web sites are differently designed. If a company wants to collect data from hundreds of sources, it is practically infeasible to try to manually write data extracting rules for each source. A fully automatic Web data extraction approaches (Crescenzi, 2001; Grigalis et al., 2012; Zhai & Liu, 2006; Zhao et al., 2005) should be used to achieve cost-effective results, and this field of automatic Web data extraction is still under active research. Even if there is a limited number of Web pages to be monitored for data extraction, a business intelligence system must have a *high degree of automation* by reducing the human efforts to build extraction rules as much as possible (Ferrara et al., 2012). In the best scenario, a user without any programming knowledge should be able to easily include a new Web site to be monitored or to fix an old one. Furthermore, online business intelligence systems extracting data from many sources should also automatically adapt data extraction rules to constant changes of Web sites.

4. A review of the state-of-the-art Web data extraction systems

In this section of the paper, we review the state-of-the-art Web data extraction systems and technological advances that have a potential to be soon integrated into online business intelligence solutions. In the first part of this section, we concentrate on reviewing commercially available of-the-shelf products. For all other systems and techniques that are not available as products, we refer a reader to a comprehensive survey by Ferrara et al. (2012). We emphasize the ease of use and strongly believe that a common user without any computer programming knowledge should be able to utilize a Web

data extraction system to extract data from simple Web pages.

4.1. Web Data Extraction Systems

Baumgartner et al. (2009) define a Web data extraction system as "a software system that automatically and repeatedly extracts data from web pages with changing content and delivers the extracted data to a database or some other application". They further divide Web data extraction tasks into five functions:

1) Web site interaction, which includes mainly the navigation to usually predetermined target Web pages containing the desired data;
2) support for wrapper generation and execution, where a wrapper is a program that identifies the desired data on target pages, extracts the data and transforms it into a structured format;
3) scheduling, which allows repeating data extracting tasks by constantly revisiting target Web pages;
4) data transformation, which includes filtering, transforming, refining, and integrating data extracted from one or more sources and structuring the result according to a desired output format (usually XML or relational database tables);
5) data provision, which is delivering the extracted structured data to external applications such as database management systems, data warehouses, business intelligence systems, decision support systems, etc.

In Fig. 1, the architecture of a typical state-of-the-art Web data extraction system is presented. The wrapper generator helps the user to build data extraction rules. It usually has a visual interface displaying the rendered target Web pages, and the user is asked to visually mark which data in a Web page should be extracted. The subunit that automatically generates the wrapper (data extraction rules) is referred to as the program generator. This module interprets the user actions on the example Web pages and successively generates the wrapper. The navigation or a Deep Web form submission in a target Web site can be recorded and later automatically reproduced. The wrapper runs previously generated wrappers stored in the

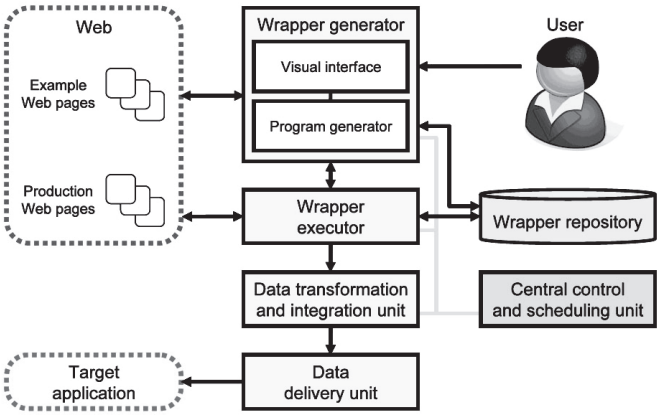

Fig. 1. ***Architecture of a typical state-of-the-art Web data extraction system*** *(Baumgartner, Gatterbauer, et al., 2009)*

wrapper repository. The data transformation and integration unit cleans, combines, transforms and integrates the extracted data. The data delivery unit delivers data via appropriate channels such as FTP, HTTP, E-mail, etc.

Although scientific literature presents numerous approaches and systems (Crescenzi, 2001; Grigalis et al., 2012; Zhai & Liu, 2006; Zhao et al., 2005) to extract data from Web pages, but only a few of these techniques are built into commercially available products. One of the most prominent examples of such systems coming from the academic research field is Lixto (Baumgartner et al., 2009). With *the Lixto Visual Developer* software, wrappers are created in an entirely visual and interactive fashion. In Fig. 2, a screenshot of the *Lixto Visual Developer* user interface is presented. In the middle, there is a fully functional Mozilla Web browser with a loaded target Web page. On the left, the navigation steps are recorded. These steps may include submitting a form, clicking on menu items, following page navigation, etc. In the bottom, there are options to configure extraction rules.

Lixto is a typical visually aided state-of-the-art Web data extraction system in which the user is asked to simply visually select the data that should be extracted. Usually, no programming knowledge is required. There are many other similar commercially available systems, that are listed in Table 1. It is worth noting that some of these systems, including the latter example of Lixto, are targeted at business customers and are not available for immediate download as an of-the-shelf product. Fortunately, the products of Mozenda, iMacros and Visual Web Ripper and some others are available to immediate download and use. Table 1 presents a complete list of the state-of-the-art Web data extraction systems. Together with names and home Web site addresses, we also specify the form in which a system is available, i.e. as a service or a standalone product. We also note whether there are a direct download link and the price.

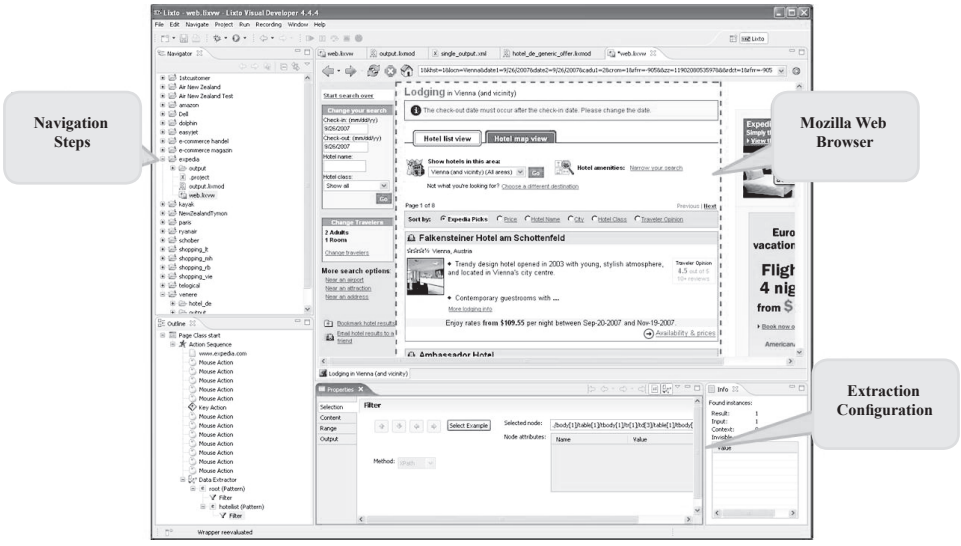

Fig. 2. ***A visual interface of the commercially available Lixto Web data extraction system*** *(Kannan, 2010)*

Table 1. ***Commercially available state-of-the-art Web data extraction systems***

Nr.	*Name*	*Form of the system*	*Available for immediate download*	*Website*	*Price*
1.	Automation Anywhere	Product	Yes	www.automationanywhere.com	From $955
2.	Connotate	Service	No	www.connotate.com	N/A
3.	Denodo	Service	No	www.denodo.com	N/A
4.	Djuggler	Product	Yes	www.djuggler.com	From $249
5.	iOpus iMacros	Product	Yes	www.iopus.com	From $495
6.	Kapow	Service	Yes	www.kapowsoftware.com	
7.	Lixto	Service	No	www.lixto.com	N/A
8.	Mozenda	Service	Yes	www.mozenda.com	From $99/ month
9.	Visual Web Ripper	Product	Yes	www.visualwebripper.com	From $299
10.	WebSundew	Product	Yes	www.websundew.com	From $69

4.2. Promising Technological Advances

Currently, commercially available of-the-shelf Web data extraction systems are best suited for extracting data from a limited number of Web sites. Even though the process of adding an additional target Web site does not usually require computer programming knowledge and is visually aided, it still needs human efforts and takes some time. Companies always search for the ways to reduce operating costs and for fully automated Web data extraction systems (Crescenzi, 2001; Grigalis et al., 2012; Zhai & Liu, 2006; Zhao et al., 2005) which do not require any human effort to define data extraction rules for any new Web site and may be a good choice. Unfortunately, we are not aware of any of such fully automatic systems to be currently commercially available as an off-the-shelf product. However, in this part of the paper, we introduce some promising technological advances in the automatic Web data extraction field, which can be soon matured into a product or could be included into current systems as a time--saving solution.

One of the most promising advances is the DIADEM project[1] developed at Oxford University. The acronym stands for Domain-centric, Intelligent, and Automated Data Extraction Methodology (Furche et al., 2012). The project is fundamentally different from many previous approaches to automatically extract Web data on a large scale. The integration of the state-of-the-art technology with reasoning using high-level expert knowledge at the scale envisaged by the DIADEM team has not yet been attempted and has a chance to become the leading example of the next generation Web data extraction systems. Without human supervision the DIADEM

[1] http://diadem.cs.ox.ac.uk/

systems locate, navigate, and analyse websites of a specific domain (such as cars, books, products, real estate, flights, etc.) and extract all contained structured data objects using highly efficient, scalable, automatically generated wrappers. The Web page analysis is parameterized using the domain knowledge that allows the DIADEM to replace human annotators and to refine and verify the generated wrappers. This approach works, in contrast to other modern Web data extraction systems which require human annotators to manually mark data on target Web pages and to record the Web site navigation. However, there seem to be remaining two challenges: simplifying the process of domain knowledge creation and adding the support for other languages. Currently, the DIADEM works only with Web sites in the English language. We see it as a limitation, since there may be many customers interested in extracting data from non-English Web sites.

Crescenzi et al. (2013) propose to leverage the power of the crowd to overcome inherited scalability limitation in data extracting systems where wrappers are generated from manually labelled examples. Crowd sourcing platforms, such as Amazon Mechanical Turk[2], present an opportunity to make the manual annotation process more affordable, also at a large scale. So, Crescenzi et al. introduce "a framework to support a wrapper inference system supervised by the crowd". Their framework aims at catching the opportunities of crowd sourcing, i.e. reducing wrapper creation costs and scaling the overall work.

Similarly, Google runs the experimental Fusion Tables project[3] which encourages creating, improving, and sharing data presented in tables. This approach, too, tries to exploit the power of the crowd in a way that many users collaboratively collect and share data from many Web sources (Gonzales et al., 2010).

Bohannon et al. (2012) from Yahoo develop a system tuned for automatic Web-scale information extraction. Like in the DIADEM project, the system is *domain-centric*. This means that there is a human supervision at the domain level, i.e. humans should define some rules concerning the topic of interest. The authors present an example that if we are interested in constructing a database of restaurants from the Web, we can specify the set of attributes that we are interested in, e.g., "name", "address", and "reviews", supply sample dictionaries or regular expressions, or language models for attributes, specify domain knowledge like "businesses typically have a single phone number but multiple reviews", and so on. Bohannon et al. believe that the domain-centric extraction can provide a promising stepping stone towards cracking the grand challenge of a general Web-scale information extraction.

5. Conclusions

The Online Business intelligence, which includes extracting, integrating, analyzing, and distributing information about products, customers, competitors, etc., helps companies to make better decisions and achieve a competitive advantage. However, extracting data from many Web sources is a human-effort-requiring process that brings considerable costs to the organization. Many promising technological advances and research directions are trying to fully automatize Web data extraction and thus help organizations to minimize the costs. However, none of the currently proposed

[2] http://www.mturk.com

[3] http://tables.googlelabs.com/

automatic Web data extraction systems are mature enough to be available as an off-the-shelf commercial product. Online business intelligence specialists in organizations can currently rely on visually aided Web data extraction systems which minimize the required human efforts to extract data from target Web sites. We further encourage researchers and industry professionals to adapt and implement in online business intelligence solutions the promising advances of automated Web data extraction technologies, such as those developed by the DIADEM team at Oxford.

REFERENCES

BAUMGARTNER, R.; GATTERBAUER, W.; GOTTLOB, G. (2009). Web data extraction system. In *Encyclopedia of Database Systems*, p. 3465–3471. ISBN 9780387355443.

BAUMGARTNER, R.; GOTTLOB, G.; HERZOG, M. (2009). Scalable web data extraction for online market intelligence. *Proceedings of the VLDB*, p. 1512–1523.

BERGMAN, M. K. (2001). The deep web: Surfacing hidden value. In *Journal of Electronic Publishing*, Vol. 7, No. 1, p. 1–17.

BOHANNON, P.; DALVI, N.; FILMUS, Y. (2012). Automatic web-scale information extraction. *Proceedings of the ACM SIGMOD ICMD*, p. 609–612.

CAFARELLA, M. J.; HALEVY, A.; MADHAVAN, J. (2011). Structured data on the web. *Communications of the ACM*, Vol. 54, No. 2, p. 72–79.

CAFARELLA, M. J.; HALEVY, A.; WANG, Z. D.; WU, E. (2008). WebTables: Exploring the power of tables on the Web. *Proceedings of the VLDB*, p. 538–549.

Connotate (2012). *Web Data Collection Monitoring Solutions.* Retrieved from: <http://www.connotate.com/solutions>.

CRESCENZI, V. (2001). RoadRunner: Towards automatic data extraction from large Web sites. *Proceedings of the VLDB*, p. 109–118.

CRESCENZI, V., MERIALDO, P.; QIU, D.; INGEGNERIA, D.; ROMA, S. (2013). A framework for learning Web wrappers from the crowd. *Proceedings of WWW*, p. 261–271.

ELMELEEGY, H.; MADHAVAN, J.; HALEVY, A. (2011). Harvesting relational tables from lists on the web. *The VLDB Journal*, Vol. 20, No. 2, p. 209–226.

FERRARA, E.; MEO, P. D. E.; FIUMARA, G.; BAUMGARTNER, R. (2012). Web Data Extraction, Applications and Techniques: A Survey. *arXiv*, 1207(0246), p. 1–48.

FLEISHER, C. S.; BENSOUSSAN, B. E. (2003). *Strategic and Competitive Analysis:* Methods and Techniques for Analyzing Business Competition. New York: Prentice Hall. ISBN 9780130888525.

FURCHE, T.; GOTTLOB, G.; GRASSO, G.; GUNES, Ö.; GUO, X.; KRAVCHENKO, A.; WANG, C. (2012). DIADEM: Domain-centric, intelligent, automated data extraction methodology categories and subject descriptors. *Proceedings of WWW*, p. 267–270.

GONZALEZ, H.; HALEVY, A.; JENSEN, C. (2010). Google fusion tables: Web–centered data management and collaboration. In *Proceedings of the ACM SIGMOD ICDM*, p. 1061–1066.

GRIGALIS, T.; RADVILAVIČIUS, L., ČENYS, A.; GORDEVIČIUS, J. (2012). Clustering visually similar web page elements for structured web data extraction. *Web Engineering*, p. 435–438.

JUNTARUNG, N.; USSAHAWANITCHAKIT, P. (2008). Knowledge management capability, market intelligence, and performance: An empirical investigation of electronic businesses in Thailand. *International Journal of Business Research*, Vol. 8, No. 3, p. 69–80.

KANNAN, N. (2010). *Online Price Intelligence for companies with real-time Changes*. Retireved from: <http://www.ebizq.net/blogs/nari/2010/05/online_price_intelligence_for.php>.

LÖNNQVIST, A.; PIRTTIMÄKI, V. (2006). The measurement of business intelligence. *Information Systems Managemen*t, Vol. 23, No. 1, p. 32–40.

MADHAVAN, J.; HALEVY, A. (2009). Harnessing the deep Web: Present and future. *Proceedings of CIDR*.

MADHAVAN, J.; JEFFERY, S. R.; COHEN, S.; DONG, X. L.; KO, D.; YU, C.; HALEVY, A. (2007). Web-scale data integration: You can only afford to pay as you go. *Proceedings of CIDR*, p. 342–350.

PISA, U.; INFORMATICA, D.; SIGNORINI, A. (2005). The indexable web is more than 11.5 billion pages. *Proceedings of WWW*, p. 902–903.

ZHAI, Y.; LIU, B. (2006). Structured data extraction from the Web based on partial tree alignment. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 12, p. 1614–1628.

ZHAO, H.; MENG, W.; WU, Z.; RAGHAVAN, V. (2005). Fully automatic wrapper generation for search engines. *Proceedings of WWW*, p. 66–75.

ŠIUOLAIKINĖS IŠ TINKLALAPIŲ DUOMENIS RENKANČIOS IR VERSLO ANALITIKAI TINKAMOS SISTEMOS

Tomas Grigalis, Antanas Čenys

S a n t r a u k a

Šiuolaikinės verslo organizacijos sėkmė priklauso nuo sugebėjimo atitinkamai reaguoti į nuolat besikeičiančią konkurencinę aplinką. Internete veikiančios verslo analitinės sistemos pagrindinis tikslas yra rinkti vertingą informaciją iš daugybės skirtingų internetinių šaltinių ir tokiu būdu padėti verslo organizacijai priimti tinkamus sprendimus ir įgyti konkurencinį pranašumą. Tačiau informacijos rinkimas iš internetinių šaltinių yra sudėtinga problema, kai informaciją renkančios sistemos turi gerai veikti su itin technologiškai sudėtingais tinklalapiais. Šiame straipsnyje verslo analitikos kontekste apžvelgiamos pažangiausios internetinių duomenų rinkimo sistemos. Taip pat pristatomi konkretūs scenarijai, kai duomenų rinkimo sistemos gali padėti verslo analitikai. Straipsnio pabaigoje autoriai aptaria pastarųjų metų technologinius pasiekimus, kurie turi potencialą tapti visiškai automatinėmis internetinių duomenų rinkimo sistemomis ir dar labiau patobulinti verslo analitiką bei gerokai sumažinti jos išlaidas.

Įteikta 2013 m. liepos 26 d.