Data Extraction for Decision-Support Systems: Application in Labour Market Monitoring and Analysis

Maxim Bakaev and Tatiana Avdeenko

***Abstract*—The paper is dedicated to the application of web data extraction technologies to facilitate decision support. As the Internet becomes a scrapable representation of the real world, the need to utilize online data arises in mafny specialized fields that search engines generally won't cover. The considerations on feasibility of employing automated web harvesting methods for obtaining data to populate decision-support systems' databases are provided. Although the initial effort required for their implementation may be substantial, for certain domains they allow getting hold of information that otherwise would be too extensive for manual collection. We also provide a description of a project involving the construction of the system used for labour market management by regional authorities. The daily data extraction from job-related websites allows unprecedented precision of analysis and decision-making in the domain.**

***Index Terms*—Artificial intelligence, data mining, information systems, web scraping.**

I. Introduction

According to estimations provided by IBM and Intel, about $2.5*10^{18}$ bytes of data is created worldwide daily, while about 10^{18} bytes of global IP data is transferred during the same time. Indeed, the World Wide Web may represent a classical case of transition of quantity into quality, i.e. high volumes of data that become available in electronic, processable, form, are to affect people's understanding of the real world.

Actually, statistical data are long-established basis for decision-making in economic behaviour, management, planning, governing, etc. However, in the digital age the traditional methods of gathering and providing statistics have some disadvantages:

- They often lag behind in time, so the data don't arrive when they're the most needed. Recently, though, there are some improvements, as statistical volumes are increasingly published online. However, the cutting of the time it takes for the information to arrive from its originating sources (companies, government institutions, etc.) definitely has its limits.
- The statistical institutions are understandably conservative and generally don't seek to embrace recent hot trends, which limits their coverage of certain fields. Their established methodologies also can't be modified overnight, so the desired arrangement of the data may be unavailable.
- The validity of the data gathered may be violated, as enterprises have low priority for submitting statistical forms. The penalties for providing distorted data are generally minimal, and the enforcement capabilities of statistical offices are nowhere near, e.g., the ones of tax administration.

Consequently, Internet is increasingly viewed as a data source [1], and even some National statistical institutes, like the one of Netherlands [2], initiate projects to augment their data. The organizations of the private sector are mostly concerned with marketing or PR-related online data, gathering information on their potential customers and their preferences, as well as monitoring social networks and the Web for opinions on certain brands or companies (see review in [3]). Not to mention the field that gave start to Internet in the first place– scientific research and development – where online citation and publications indexing services have been dominating for more than a decade.

Naturally, the vast amounts of data in most cases call for their automated processing, involving artificial intelligence (AI) methods and intellectual information systems. These applications include:

- Automated control and management – e.g. based on data from enterprise information systems or stock exchange quotes.
- Data mining and information extraction with natural language processing – e.g. analyzing news, company or brand references online, etc.
- Knowledge engineering, in both human-readable and machine-readable forms, and automatic inference.
- Computer-aided design and decision-making support, which probably imply the most intense involvement of human, being rather intelligent-amplifying than artificially intelligent.

There is actually a belief in the coming of Semantic Web that would make all online data structured, so that one would be able to run queries directly on them. Still, despite the unceasing interest towards this topic, the Semantic Web seems not to be overtaking just yet. For now, it's reasonable to assume that the majority of information on the Internet is there for humans, not computers, and remains largely unstructured. Thus, data extraction methods must cope with the current situation, and the following issues can be identified with it:

- Legal issues – it is generally considered legal to collect information that is openly available online, unless the source explicitly forbids doing so. In some countries the legislation may be different, but this is mostly beyond

Manuscript received June 15, 2013; revised August 9, 2013. This work was supported in part by the innovative young scientists' grant of the City Hall of Novosibirsk, Russia: "The system for automated monitoring and intellectual analysis of labour market".

M. Bakaev and T. Avdeenko are with the Economic Informatics department, Novosibirsk State Technical University, Novosibirsk, 630132, Russia (e-mail: maxis81@gmail.com, tavdeenko@mail.ru).

the scope of our paper.

- Technological issues – there's no single universally accepted and replicable solution at the moment, as research and development in this field goes on. The global search engines are focusing on unstructured textual data, while individual "web scraping" tools are context-dependent or require a considerable effort to "learn" how to do data extraction from particular sources. So, technological boundaries and solutions will be given special consideration further.
- Availability issues – one may have noted the disparity between the amounts of new data created and data transferred that we specified in the beginning of the paper. We are to provide some reflections on what does and what does not have the "online footprint".
- Efficiency issues – if online data gathering is to replace or augment traditional methods, it must have advantages, such as lowering costs or providing benefits and opportunities in terms of better speed, frequency, quality, etc. This is covered quite well in [2], and we will provide some insights and reflection.

So, overall, our paper is dedicated to feasibility of online data extraction to support decision-making in governing, management, etc. In Section II we provide some overview of the state-of-art in the field, while in Section III we describe the real case – the creation and utilization of software system for scraping web data related to a regional labour market and aiding local authorities in managing it.

II. Data Extraction for Decision-Support Systems

The classification of sources for data extraction naturally follows the classical "sender – channel – receiver" triad. Consequently, the three major categories for data extraction methods are [1]:

- User-centric – aimed on the data that originate at an individual's client, such as PC or mobile device. This may be a powerful tool in gathering statistics on users' behaviour, including online spending, geo location, etc., though it generally requires the most processing power to aggregate data. However, individual users are likely to object to their data being collected, as the experience of mobile applications developers shows us, and obtaining their consent may be the most vital matter. If no respective software (even if it isn't directly installed by the owner), exists at the client, no data collection will occur.
- Network-centric – focused on data in transit, i.e. flowing in a network. Again, legal and social issues are very intense here, as deep packet inspection is currently banned in most cases. That basically leaves such attributes as sender and receiver types, the volume of data being transferred, and protocol type. Thus, it may be probably stated that this group of methods' applicability is quite narrow, although this may change in the future.
- Site-centric – dealing with data that originate at web servers. This may involve extracting data from web pages or feeds output by the server, receiving data directly from the server's storage, such as through database API, or examining attributes of the site itself, e.g. technologies and platforms used, quality and quantity of web pages, response times, etc. This is the most established and long-standing group of methods, as search engines actually fall into this category. However, website data extraction is "supply-driven", so a major problem is discovering how a decision can be supported by existing websites and available data.

The following chapters discuss in more details the issues of technology, availability, and efficiency specified above for site-centric data extraction methods.

A. Technological Issues

In terms of computational complexity, the problem of structured information extraction from web pages is said to be polynomial or even linear for specific domains [4], although it may not be the same for processing this information, especially if it involves natural language processing. Still, it can probably be assumed that it's theoretically possible to build sufficiently prompt decision-support systems operating on online data, even for large amounts of them. So the major concerns in data and information extraction pertain to a different area and are reflected in popular criteria of accuracy and completeness [5]. The former is generally understood as the ratio of objects correctly extracted by method to some "true" number of objects, e.g. the one that would be extracted by human expert (there are also estimations that a web page contains about 5000 elements on average [4]). The latter is more delicate, being the ratio between "correctly" identified objects to the total number of objects extracted.

The major threats to completeness and accuracy in data extraction from websites are the sites' complexity and changeability. By complexity we imply the inevitable utilization of technologies beyond plain HTML – the construction of queries to database, information disclosure with AJAX and its storage in files (pdf, doc), usage of Flash, etc. Changeability is also an intrinsic feature of websites, meaning that the structure of sites and web pages change with time, trying to put up more data, improve interface and so on. The up-to-date technological apparatus accommodating these issues is the following.

About a decade ago, quite an important innovation in the field was put forward – the automated generation of *wrappers* (see the definition and detailed review in [6]), so that an algorithm for structured scraping a particular webpage data is created automatically, based on the page contents' analysis. This solution has a potential to significantly reduce the programming effort associated with starting data extraction from a new website, but in most real cases interactive involvement of a human operator remains advisable. The approaches for wrapper generation include regular-expression-based, logic-based that employ special wrapper programming languages, and tree-based that split web page into so-called data record regions. The AI field goes further in this respect, aiming to develop solutions that would learn rules for extracting information, often also under a human supervision initially (see review in [6]). It is believed that unsupervised data extraction can be achieved with the use of ontologies, which however have to be

domain-specific and may require even more considerable effort for building and maintaining [7].

At the same time, complexity remains a challenge for all the above approaches, as sometimes it prevents wrappers from accessing all the data available on web pages in the way a human user would. The recent spread of AJAX-based technologies had especially highlighted this problem, although there are solutions that more or less successfully emulate human user behaviour in web browser or employ a special script-executing machine (an overview of the state of art is provided in [8]). Quite a novel idea here is to take a web page just as a human would, and rely on image recognition techniques to extract data from it. There are evidences, e.g. with ViDE approach, that visual wrappers generation indeed yields solid results [9].

All in all, the technological issues of automated data collection from websites seem to be more or less resolved (except possibly for Flash-based data), although quite advanced methods and tools are often required. Still, many algorithms and products are available on open-source of free-to-use basis, such as RoadRunner, or XPath, while there are already business data collection software-as-service solutions such is the renowned Mozenda or Diffbot.

B. Availability and Efficiency Issues

The following are considered to be among the advantages of automated online data gathering compared to traditional, manual one [1], [2]:

- Increased efficiency in terms of labour-intensiveness, for both gatherers and possible respondents. This also may mean lower costs, though it's not always the case.
- Extensiveness of dataset, as higher efficiency allows recording more attributes of studied objects, increase the sample size, etc. Still, it shouldn't substitute the quality, and the problem of defining meaningful indexes for monitoring remains.
- Timeliness and frequency, which mean that data can be gathered in real-time and with unprecedented regularity, if the studied field calls for it. This is often helpful in studying short-lived phenomena, especially the ones originating in network environment. At the same time, the data collection must be prompt and flexible enough to cope with the transience of the object.
- Better validity of data due to the removal of respondent burden (as information is gathered indirectly) and manual errors prevention. However, this is only true with the assumption that suitable data is available and the scraping algorithms work correctly.

The costs involved with automatic data gathering are indeed not necessarily dramatically lower than for manual work. Third-parties that provide relevant solutions seek to maximize their returns, while in-house development for a non-specialized company may be understandably costly. Also, there's always the cost of extraction customization, even if not development, for a particular case, as well as subsequent maintenance – practical experience shows that break-even point is highly sensitive to sites' changeability and the data extraction robots re-programming costs [2]. It seems that currently, despite admitted advances in automated wrapper generation, human involvement should be still required to monitor the data extraction in long-term.

Extensive and frequent collection of data can't make up for its validity and relevance, so consideration of methodological issues is still very important in automated data extraction. For online data extraction, the distinction between the *unit of analysis* – the studied object – and the *unit of observation* – technical data on some properties of the object – is particularly prominent [1], which requires careful selection of meaningful indexes. Potential threats to validity are non-random sampling and poor generalization, as information available online is, well, only representative of online universe.

On the whole, the availability of data related to the unit of analysis depends on whether there are any *online information transactions* that involve it. Naturally, the *transaction* means the existence of both information source and recipient – as experience shows, if the second side in the potential transaction is unclear (i.e. there's no established target user), the data won't be generated or updated in the long run. *Online* is also an important clause here, as significant amount of information never leaves internal databases of companies or government institutions, being unapproachable for collection.

The positive examples include stock and currency markets, real estate ads, labour markets, prices for goods and services, etc., so clearly economic incentives play a major role here. Among possible sources for extraction, "flock points" that draw companies or individuals to voluntary put up their data for information exchange are generally preferred. Indeed, if the necessary data are scattered among many diverse websites, the number of different scraping algorithms may become too high, imposing prohibitive development, customization or maintenance costs. The same reason applies to favor sources where the unstructured data is still more or less organized, with proper HTML (XML) mark-up and meaningful CSS styles.

Finally, we would propose the following steps for rough estimation of automated online data extraction feasibility:

1) Does a *customer* exist for the data collection? Are there any decisions to be made based on the data? Or is there, at least, any added value associated with aggregated data?
2) Are there clearly definable units of analysis and do their attributes manifest themselves in online information transactions, providing a unit of observation? What should be the measurement indexes?
3) Is it legally and socially acceptable to collect these kind of data?
4) What are the sources (websites) for data extraction and what is the number of them that would sufficiently represent the universe? What is the validity of data and what generalizations can be made?
5) What kind of sources are these: is their output properly organized, do they use any technologies that will complicate data extraction, what is their expected changeability rate and will the changes significantly affect the output structure? Do these sources and their information providers have sound reasons to last?
6) What would be the desired and the doable frequency of

data collection? Will increased frequency, better timeliness and greater extensiveness of data afford a significant improvement over manual collection?

By answering the above questions, one may estimate feasibility of initiating an automated data extraction project, although it surely doesn't eliminate the need for careful calculation of benefits and costs. Summarizing, we may say that the most promising application of automated data collection is not in replacing existing manual procedures, but in reaching new areas and achieving higher level of detail. There are decisions to be made in domains where the amount of data generated or updated daily is beyond any processing that can be done by hand, so automation and intellectualization are the only reasonable options. In the following section of our paper we describe a real case of developing data extraction intellectual system for Novosibirsk City Hall's department responsible for labour market management.

III. Application in Labour Market Monitoring and Analysis

The developed software system is dedicated to supporting decision-making in labour market management by the City Hall of Novosibirsk, Russia. The following are the answers to the feasibility-estimation questions that we proposed above:

1) The customer is the Labour Committee of the City Hall, which needs labour market-related information to perform its authorized functions.
2) The major units of analysis are *job positions* (vacancies) in companies and *job applications* (resumes) of individuals. There are indeed online information transactions between hiring companies and job-seekers, the major measurement indexes being wages proposed and wages requested, as well as relation between the number of vacancies and resumes. However, data on the currently employed workers won't be available, as it largely remains internal.
3) These data are voluntarily put online by respective parties, so its collection is both legally and socially acceptable, especially for aggregation for a government institution.
4) There are dedicated websites for posting vacancies and resumes. The number of federal and regional websites that would cover more than 90% of Novosibirsk's online job-related ads is no more than 5. The available data on the wages is valid in terms of demand and supply. Generalizations can be made for relatively advanced industries, as e.g. for agriculture the number of ads will be minimal.
5) The sources in consideration provide properly organized output in HTML, although they occasionally start using AJAX technologies which may complicate extraction. For each of them, minor changes in structure and information output are likely to occur once every three moths, and major ones – once every 18 months. Most of these websites are more than 10 years old, have established business models in providing the platform for job-related ads and are likely to remain operational for the next several years.
6) Data extraction should be performed daily, as many ads are short-lived, and it seems technologically possible. The traditional, off-line ways of gathering respective data are incomplete or too infrequent and costly. The dedicated governmental organization, the Centre for People's Employment, by and large works with citizens who seek to obtain unemployed status, so it alone couldn't provide valid information on labour resources demand or supply in the city.

In the result, we developed and configured the software system able to automatically collect data from specified online sources that openly and massively publish job-related ads. The system has three major tiers:

- The data gathering module, responsible for accessing the source websites and grabbing data from webpages. Technical data are removed from HTML (to optimize the system's database for storage space), but otherwise the entire webpage's code is saved. A sample regular expressions-based algorithm (configurable) for data extraction from a web source is provided in Appendix.
- The processing module, responsible for structuring the data. Currently it extracts information related to vacancies and resumes and their specified properties, as well as divides the jobs by industries (sectors) – with some simple natural language processing.
- The analysis module, directly responsible for decision-making support and providing capabilities for reports generation, filtering, notifications, etc.

The data collection started in late 2011 and by the summer of 2013 the system's database contains about 800 thousand analyzed records and amounts to 4 GB. About 1000 vacancies and 500 resumes are added daily, and it only takes the system about 1 hour to extract and process these data, so a significant reserve still remains, although the designers initially expected higher load. Fig. 1 shows an example report for "Information and communication technology" sector for the last year – average wages proposed by companies and monthly number of vacancies. For this sector, the system also uncovered an interesting fact that wages proposed (avg. 44.5 thousand roubles/month) are much higher than the wages asked for (avg. 25.9 thousand roubles/month), which emphasized the lack of highly-qualified IT specialists on the free market – most of the resumes belong to students, interns or freelancers.

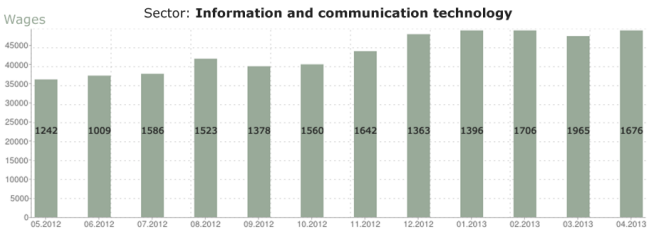

Fig. 1. Sample report for "Information and communication technology" sector (monthly wages as bars and monthly vacancies as numbers).

IV. Conclusions

In our paper we discussed some issues related to automated online data extraction to support decision-making process. We provided an overview of the field, highlighting

technological aspects of web data scraping and giving some insights on when it can and should be performed. We also proposed a qualitative method for estimating feasibility of employing automated collection of data from websites. A description of a real case is provided, which is the development of software system for labour market monitoring and analysis, used by the City Hall of Novosibirsk, Russia.

APPENDIX

Source name: *hostname1.ru*
Source encoding: UTF-8
URL:
vacancy%http://*hostname1.ru*/vacancy?search_key=d88912&limit=100&order_by[]=orderby_date&order_dir[]=desc
resume%http://*hostname1.ru*/resume?search_key=40cveym&limit=100&order_by[]=orderby_date&order_dir[]=desc
Sectors:
vacancy%http://*hostname1.ru*/resume/add/?unregistered|select#select_rubric option
resume%http://*hostname1.ru*/resume/add/?unregistered|select#select_rubric option
Block with page-based navigation:
vacancy%div.ra-element-paginator-holder ul.ra-pagination-pages li.ra-pagination-item span
resume%div.ra-element-paginator-holder ul.ra-pagination-pages li.ra-pagination-item span
Variable for page-based navigation:
vacancy%page|1 resume%page|1
Block with hyperlink:
vacancy%div.ra-elements-list div.ra-elements-list__item a.ra-elements-list__title__link resume%div.ra-elements-list div.ra-elements-list__item a.ra-elements-list__title__link
Filter for hyperlink:
vacancy%/vacancy/* resume%/resume/*
Publication date:
vacancy%div.ra-elements-list div.ra-elements-list__item div.ra-elements-list__date resume%div.ra-elements-list div.ra-elements-list__item div.ra-elements-list__date
Block with content:
vacancy%#ra-vacancy-full resume%div.view-position
Removed blocks:
div.specials-list|span.specials-list-item-short div.toolbar-post div.noprint span.num-posts div.photo-person-right a#reformal_tab div.ra-vacancy-full-info-data-post
== Extracted data ==
Name:
vacancy%div.ra-vacancy-full-title h2 resume%h3
Sector:
vacancy%.ra-vacancy-full-see-also-column-item-title a|0 resume%span.specials-list-item-short|0
Other data:
vacancy%wage|str|div.ra-vacancy-full-salary|0 vacancy%company|str|div.ra-vacancy-full-flying-box-title|0 resume%wage|str|div.pay|0

REFERENCES

[1] European Commission, "Internet as data source. Feasibility Study on Statistical Methods on Internet as a Source of Data Gathering," Report for the European Commission DG Communications Networks, Content & Technology, Dialogic, Netherlands, 2012, pp. 243.

[2] R. Hoekstra, O. Bosch, and F. Harteveld, "Automated data collection from web sources for official statistics: First experiences," *Journal of the International Association for Official Statistics*, vol. 28, no. 3-4, pp. 99-111, 2012.

[3] G. Erétéo, M. Buffa, F. L. Gandon, P. Grohan, M. Leitzelman, and P. Sander, "A state of the art on social network analysis and its applications on a semantic web," in *Proc. SDOW Workshop of 7th International Semantic Web Conference (ISCW 2008)*, Karlsruhe, Germany, 2008, pp. 6.

[4] B. Kraychev and I. Koychev, "Computationally effective algorithm for information extraction and online review mining," in *Proc. the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS'12)*, Craiova, Romania, 2012, art. 64.

[5] Y. A. Zagorulko and E. A. Sidorova, "Document analysis technology in information systems for supporting research and commercial activities," *Journal Optoelectronics, Instrumentation and Data Processing*, vol. 45, no. 6, pp. 520-525, 2009.

[6] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner. Web data extraction, applications and techniques: a survey, *ArXiv preprint* [Online]. pp. 1-40. Available: http://arxiv.org/abs/1207.0246

[7] S. Flesca, T. Furche, and L. Oro, "Reasoning and ontologies in data extraction," *Reasoning Web. Semantic Technologies for Advanced Query Answering. Lecture Notes in Computer Science*, vol. 7487, pp. 184-210, 2012.

[8] S. Choudhary, M. E. Dincturk, S. M. Mirtaheri, A. Moosavi, G. von Bochmann, G. V. Jourdan, and I. V. Onut, "Crawling rich internet applications: the state of the art," in *Proc. the 2012 Conference of the Center for Advanced Studies on Collaborative Research*, November 2012, pp. 146-160.

[9] W. Liu, X. Meng, and W. Meng, "ViDE: A vision-based approach for deep web data extraction", *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 447-460, 2010.

Maxim Bakaev graduated from Novosibirsk State Technical University (NSTU, Novosibirsk, Russia), as Master of Information Systems in Economy. He is also Master of Digital Design, Graduate School of Digital Design Kyungsung University, South Korea, 2007. In 2012 he received his PhD in Software Engineering from NSTU with the dissertation "The Development of Intellectual System Supporting Human-Computer Interaction in Web Applications".

He is a senior assistant professor of Economic Informatics Department of NSTU and Vice-Dean of Business Faculty. His research interests include Human-Computer Interaction, Interface Design and Usability, Knowledge Engineering, Universal Accessibility.

Maxim Bakaev is a member of International Association of Computer Science and Information Technology (IACSIT) and reviewer for various international conferences, including KEER and NordiCHI.

Tatiana Avdeenko graduated from Novosibirsk Electrical Engineering Institute (now Novosibirsk State Technical University – NSTU, Novosibirsk, Russia), Department of Applied Mathematics. In 2004 she received Doctoral degree in Mathematics Modelling, Calculus of Approximations and Software Systems, awarded by NSTU.

She is a full professor and the Head of Economic Informatics Department of NSTU. She is the author of more than 90 scientific papers including 1 monograph and 5 textbooks. Her research interests include Knowledge Engineering and Management, Artificial Intelligence, Information and Decision-Support Systems.

Tatiana Avdeenko is a Certified Expert of the Russian National Agency for Accreditation in Education, and reviewer for various international conferences, as well as for the journal Inverse Problems in Science and Engineering. She has The Honour Letter from the Ministry of Education and Science of Russian Federation.