Prospects and Challenges in Online Data Mining

Experiences of Three-Year Labour Market Monitoring Project

Maxim Bakaev(✉) and Tatiana Avdeenko

Economic Informatics Department,
Novosibirsk State Technical University, Novosibirsk, Russia
bakaev@corp.nstu.ru, avdeenko@fb.nstu.ru

Abstract. The paper provides reflections on feasibility of online data mining (ODM) and its employment in decision-making and control. Besides reviewing existing works in different domains of Data Mining, we also report experiences from ongoing project dedicated to monitoring labour market with the aid of dedicated intelligent information system. Benefits of ODM include high efficiency, availability of data sources, potential extensiveness of datasets, timeliness and frequency of collection, good validity. Among special considerations we highlight a need for sophisticated tools, programming and maintenance efforts, hardware and network resources, multitude and diversity of data sources, disparity between real world and Internet. Finally, we describe some examples of the intelligent system application, in particular analyzing labour market data for several regions.

Keywords: Data Mining · Intelligent information systems · Big Data · Web content mining · Decision-making

1 Introduction

As the amount of data being created and copied in the world currently has the order of 10^{21} bytes, it is no wonder that many major companies and organizations seek to intensify Big Data (BD) research and application [1]. So, BD became an established field in quite a short timeframe, and a significant share of research works is already recounting the state-of-the-art in it. For example, a review of the field progress and the relevant technologies is provided in [1], while [2] looks back on the development of Data Mining (DM), which is often viewed as most essential component for successful utilization of BD. The recitation of popular DM algorithms was already done in 2007 [3], and by now they, as well as specially adapted methods (see, e.g. [4]), are widely employed for processing BD. This prompt development is in particular stimulated by the need to support decision-making, management, and control in general, which can no longer rely on traditional statistics, such as administrative registers. Among their disadvantages that are often noted, is virtually unavoidable time lag, conservatism in

Y. Tan and Y. Shi (Eds.): DMBD 2016, LNCS 9714, pp. 15–23, 2016.
DOI: 10.1007/978-3-319-40973-3_2

terms of measuring novel concepts and phenomena, inability to embrace rapid trends, lack of coverage of certain sectors of economy, etc. [5].

At the same time, the development of DM is not uniform per different domains, and it is said to be problem-oriented [2]. Quite a lot of progress has been made recently in medicine, where advancement of expert and intelligent systems remains a mainstream [6], and it's expected that Big Medical Data classification and analysis will be able to significantly decrease the ever-growing health care expenses in developed countries. Extensive review of milestones in the field since the beginning of the millennium can be found in [7], where principal DM approaches are listed as follows: classification, regression, clustering, association and hybrid. Overall, already in this domain the desired goals seem to be agreed upon, the data collection and structuring have been carried out for a long time, and current research is to a significant degree aimed on benchmarking and perfection of various mining methods, compared per metrics of accuracy, sensitivity, specificity, etc. [6].

Another example of a long-established field is Business Intelligence and Analytics, which lately has also been seeking to harness BD – a review of research and bibliometric study can be found in [8]. The development in this domain made it clear that having enough data and decent processing algorithms are not sufficient, as most problems seem to arise due to ambiguous objectives and lack of clear indexes to be analyzed or monitored. Also, unlike in medicine and health care domain, major organizations in business may be reluctant to cooperate in research, to share data or agree on their common structure. We'd also like to highlight some relatively novel applications of DM, one of which is Educational DM, whose increasing popularity is understandable in the light of boom in online educational services. The author of [9], having considered 240 works in the field, notes that it is still incipient, as only a few options from DM repertory are used. Another new challenge in BD analysis is said to be the Internet of Things, with its naturally voluminous and diverse data that are considered "too big and too hard to be processed by the tools available today" [10, p. 78]. We'd also like to highlight that unless some aggregators are employed, these data should be available at innumerable nodes *(things)* accessible via online channels, which would make their collection quite arduous. Indeed, ODM, when no readily open databases or dedicated data outputs exist, should be viewed as a particular domain, due to specifics of data allocation, acquisition, structuring, etc. Currently, its most popular applications are rather "qualitative" ones, such as social networks mining, usually for e-commerce (see review in [11]), or semantic mining, for which numerous tools exist, such as described in [12]. However, for decision-making, management, control, etc., often quantitative analysis of online data is highly desirable, though it seems to receive less attention [13] – most of available works deal with merely technical aspects of web mining, not enquiring how information obtained online could be used and what decisions might be improved with it.

Thus, in our paper we reflect on prospects in ODM, as well as up-to-date challenges in this field – by which we mostly consider *web content mining*, not *web structure mining* or *web usage mining* (see explanation of the distinction in [14]). In Sect. 2, we provide some theoretical considerations and review existing research works. In Sect. 3 we describe an ongoing project, intelligent system that performs collection, processing, and analysis of online data related to labour market. We provide an example of real

data collected by the system and how it could be used in decision-making, control and forecasting. Overall, we would like to encourage discussion on theoretical and practical aspects of employing online data, on automation and intellectualization of this process.

2 Online Data Mining

Analyzing advantages of employing online (web) data in control and decision-making, put forward by various research works, such as [13–15], coupling them with our own practical experiences described later in the paper, we composed the following list:

- **High efficiency** in terms of labour-intensiveness, which *may* also mean lower costs. However, break-even point is highly sensitive to the number of websites, their changeability, volume and structure (if any) of collected data.
- **Open availability of data sources** – websites that voluntarily publish data to be accessed by human visitors or, in rarer cases, robots. Other factors being equal, data sources that have better coverage (that is, more data and more frequent updates) and more stable output in terms of data structure (proper HTML/XML mark-up and meaningful CSS styles) are first candidates for ODM.
- **Extensiveness of datasets**, as Internet is surely a Very Big Data source. Higher efficiency and data availability allow to dramatically increase sample size, possibly even extending the coverage to the whole population, as well as to record more attributes of studied objects. However, generalizations are to be made carefully, as information available online is only representative of online universe.
- **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 that are common in the WWW.
- **Better validity of data** due to the removal of respondent burden (as information is gathered indirectly) and prevention of manual-processing errors. However, this is only true if suitable data are available and the scraping algorithms work correctly and accurately maintained.

As for problems and special considerations of ODM, we'd like to note the following:

- **Much more sophisticated tools**, compared to general Data Mining, are necessary to collect online data, which are prone to changeability and sometimes hard-to-get. Indeed, algorithm for scraping a particular webpage data can be created automatically, based on the page contents' analysis, with the use of the so-called wrappers technology (see the definition and detailed review in [16]) that has a potential to significantly reduce programming effort. However, it seems that currently, despite admitted advances in automated wrapper generation, human involvement is still required to maintain data extraction in the long run. Thus, the availability of open-source or free-to-use components and products such as RoadRunner or XPath, or commercial solutions such as Mozenda, Diffbot, etc. (see in [16]), doesn't fully resolve the issue.

- **More hardware and network resources** needed, even though 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 [17] (there are estimations that a web page contains about 5000 elements on average). The situation is more complex for Rich Internet Applications, where both all URLs **and** all application states have to be attended by a data scrapping algorithm, so that the task can be mapped to directed graph exploration problem. Still, satisfactory performing methods exist (see review in [18]), such as distributed greedy algorithm or vision-based approach (see ViDE algorithm [19]). In should be noted that Flash-based websites currently remain largely inaccessible for automated online data mining, but their number is already comparably low.
- **Incompleteness of Internet** – not all objects of the real world have "online footprint", but only those involved in some kind of online information transactions. Naturally, the existence of economic benefit for both sides is a good motivation – well-known examples of online data suitable for mining include e-commerce prices for goods and services, stock and currency markets, tickets costs and availability, real estate, labour markets, etc. [13]
- **Multitude of sources**, which leads to increased need for efforts and resources required to collect and process data, given high probability of inconsistencies between sources. If online data are scattered among many diverse websites, the number of different scraping algorithms may become too high, imposing prohibitive development, customization or maintenance costs. In certain fields it is possible to select a number of data sources that would make up satisfactory sample, but this may lead to threats in data validity.
- **Threats to data validity**, as data sources are selected without non-random sampling, but based on their data volume, technical properties, accessibility, etc.
- **No quantity can make up for understanding the goals and means** – ODM has to start with "why" and "for whom" questions (see a simple algorithm for estimating the feasibility of online data employment in [21]), unlike general DM that often seem to commence with "we have these data, what can we do with them?".

So, with the above points we tried to justify the claim that ODM is indeed a distinct field, requiring special methods, tools, and methodological considerations. The following section of our paper shows how these were reflected in a real project – intelligent system to analyze regional labour markets, – and what new lessons we could learn.

3 The Labour Market Online Monitoring Project

3.1 The Project Background

The developed software system is dedicated to supporting decision-making in labour market management by the City Hall of Novosibirsk, Russia. The system was put into operation in 2011 and currently its database contains more than 10 million records on vacancies and resumes for Novosibirsk and certain other regions (more details can be found in [20, 21]).

The data structuring data is straightforward in most cases, as web pages content fields are marked with respective id names and the implicit data model is known to us. Often it's desirable to exclude multiple copies of the same item, e.g. collected from different websites or in different time periods, from the analysis. The first step in the intelligent filtering process is generation of hash-code for the collected page, with all the variable elements, such as counters or advertisements, removed. This approach works for most of the copies and excludes them from further structuring, thus decreasing the system load. However, for more accurate filtering on the second step, structured data are used – they are merged into a string for which we also generate a hash-code for subsequent validation and comparison. The resulting productivity, even given the high volumes of data in the system, is quite satisfactory, and the filtering mechanism can work as a daemon when the system load is low in the absence of data collection or processing.

3.2 The Project Experiences

- **Feasibility may vary.** The advantages of automated online data collection are not guaranteed, and the most important factors are: (a) whether the object of analysis has a reliable "online footprint", i.e. enough of its properties are manifest online, regularly updated by an interested party, and relatively trustworthy; (b) whether the concentration is high enough, i.e. most of the information universe is represented on a handful of websites – about 10, not 1000, which would make the system maintenance nearly impossible.
- **Changeability and monitoring.** Inevitably, the structure of sites and webpages change with time, as they try to put up more data, improve interface, introduce more advanced technologies like AJAX, and so on, which significantly affects accuracy and completeness of ODM, and introduce considerable maintenance costs. While the latter seem unavoidable, the former could theoretically be aided by a monitoring sub-system, overseeing correctness of data collection. However, setting up monitoring is not a trivial problem, as only relatively simple data gathering malfunctions can be validated, but generally human programmer involvement is still required periodically.
- **This is not Big Data.** The online data collection, in most cases, seems to be of "Small Data" domain. That is, all of the accessed data can be stored and re-processed later if necessary (e.g., with more advanced algorithms or to extract more object of analysis's properties). We even store HTML code of webpages, although with certain optimization, and still our database size that after 3.5 years of the system's operation contains about 10 million records, amounts to about 120 GB.
- **Customer doesn't know what to ask for.** Decision-making based on online data is quite a novel approach, and it was confirmed in our project that management, at least on municipal level, does not fully grasp its potential. So, the customer, for whom data is collected and analyzed, cannot directly drive the development of the system.

3.3 Highlights of Labour Market Online Data Analysis

Below we provide some data from our system for selected regions (*obl.*, *krai*) of Russia (mostly based in Siberian Federal District), gathered and processed up to the end of 2014 (as the data disclosure was approved by the system's management). Table 1 contains data on average weekly numbers of vacancies and resumes (so that only unique items are considered, not repeating postings). The official data on the regions' urban population is for the 1st of January, 2015. The data for 2015 was also collected and will be analyzed and subsequently made available with the approvement by the system's management.

Table 1. Average weekly numbers of vacancies and resumes per regions, 2014.

Region	Vacancies (weekly)		Resumes (weekly)		Ratio (V/R)
	Total	Per 100,000	Total	Per 100,000	
Krasnoyarski krai	997	45.45	639	29.13	1.56
Kemerovsk. obl.	1446	61.91	1202	51.47	1.20
Novosibirsk. obl.	3204	148.55	3300	153.00	0.97
Omskaya obl.	1102	77.19	729	51.06	1.51
Tomskaya obl.	825	106.72	659	85.25	1.25
Tyumenskaya obl.	539	18.83	444	15.51	1.21

From the data presented in Table 1, we can conclude that the considered regions significantly differ in the numbers of vacancies and resumes published online. While Novosibirsk region is special, as we noted above, the Tomskaya oblast still has notable lead from the others. It indeed can be explained by higher "online mobility" of its citizens, quite a significant share of which is students. Interestingly, the ratios between vacancies and resumes are more stable, although in most of the regions there are more companies looking for staff than people looking for jobs. We are continuously monitoring this ratio, which is a good indication of the state of economy.

Table 2 presents data on average salaries that are proposed in vacancies and requested in resumes. We compare these data to the official salary statistics (which is taken for the whole year 2014) and calculate the ration between the average salaries data from the system to the official data. The official salary data for Tyumenskaya obl. is taken excluding the ones in its autonomous regions, which are rarely present online but introduce bias due to high salaries in their oil industry. Table 3 contains the salary-related information for Novosibirsk region by years. The data is shown for half-years, but official salaries are taken for the whole year, due to lack of more detailed statistics. It's interesting to note that in 2013, when the economic situation was quite stable, the ration between proposed and requested salaries (Ratio V/R) was close to 1, but how the situation changed in 2014, when the economy worsened in the fall. Prior to the decline, the requested salaries boomed (it's a well-known fact that the cost of resources increases prior to crises), but after most of the workforce experienced the economic troubles, they got ready to work for much lower salaries.

Table 2. Average monthly salaries per regions, 2014.

Region	Salary (Rubles per month)			Ratio (V/R)	Percentage of official
	Official	In vacancies	In resumes		
Krasnoyarski krai	34224	30915	25514	1.21	82.4 %
Kemerovskaya obl.	26732	30124	22938	1.31	99.2 %
Novosibirskaya obl.	27267	33110	22100	1.50	101.2 %
Omskaya obl.	26313	27350	24538	1.11	98.6 %
Tomskaya obl.	32503	27807	24809	1.12	80.9 %
Tyumenskaya obl.	34221	38865	28566	1.36	98.5 %

Table 3. Average monthly salaries for Novosibirsk region, per years.

Period (half-year)	Salary (Rubles per month)			Ratio (V/R)	Percentage of official
	Official	In vacancies	In resumes		
2012 (I)	23246	26986	22796	1.18	107.1 %
2012 (II)	23246	27151	23973	1.13	110.0 %
2013 (I)	25528	24798	23355	1.06	94.3 %
2013 (II)	25528	25370	25774	0.98	100.2 %
2014 (I)	27267	27589	30563	0.90	106.6 %
2014 (II)	27267	33110	22100	1.50	101.2 %

4 Conclusions

Nowadays, as annual growth rate in the amounts of data created and transferred in the world is estimated as 50 %, Internet is increasingly viewed as a data source, and already not only business organizations, but even understandably conservative National statistical institutes initiate projects to employ online data. It is said that BD and DM are domain-specific [2], and in such fields as health care (medical data), business intelligence, education, etc. they are on different stages of maturity. Thus we reason that ODM should be seen as a distinct field as well, having very specific considerations in terms of data collection, methodological aspects and so on.

The paper discusses certain issues in intelligent online data employment in analysis of various phenomena and supporting decision-making. Although technological problems seem to be mostly resolved [17, 18], the matter of general feasibility seems to be less explored [13]. Apparently, 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 hand-processing, so automation and intellectualization are the only reasonable options. From our relevant experiences from a project ongoing from 2011, the intelligent information system for monitoring labor market, we in particular conclude that an important challenge in development such systems is their potential users' (that is, government agencies officials) conservatism with IT, which puts requirements generation burden on developers.

We also provide real data from the system, for selected regions of Russia and various time periods, and offer economy-related observations. Interestingly, we can conclude that capabilities of our and similar intelligent solutions so far exceed the information needs of responsible authorities, and we attempt to outline directions for securing wider acceptance of ODM methods.

Acknowledgement. This work was supported by RFBR according to the research project No. 16-37-60060 mol_a_dk.

References

1. Chen, M., Mao, S., Liu, Y.: Big data: a survey. Mob. Netw. Appl. **19**(2), 171–209 (2014)
2. Liao, S.H., Chu, P.H., Hsiao, P.Y.: Data mining techniques and applications–a decade review from 2000 to 2011. Exp. Syst. Appl. **39**(12), 11303–11311 (2012)
3. Wu, X., et al.: Top 10 algorithms in data mining. Knowl. Inf. Syst. **14**(1), 1–37 (2008)
4. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Trans. Knowl. Data Eng. **26**(1), 97–107 (2014)
5. Beręsewicz, M.: Estimating the size of the secondary real estate market based on internet data sources. Folia Oeconomica Stetinensia **14**(2), 259–269 (2014)
6. Seera, M., Lim, C.P.: A hybrid intelligent system for medical data classification. Exp. Syst. Appl. **41**(5), 2239–2249 (2014)
7. Esfandiari, N., Babavalian, M.R., Moghadam, A.M.E., Tabar, V.K.: Knowledge discovery in medicine: current issue and future trend. Exp. Syst. Appl. **41**(9), 4434–4463 (2014)
8. Chen, H., Chiang, R.H., Storey, V.C.: Business intelligence and analytics: from big data to big impact. MIS Q. **36**(4), 1165–1188 (2012)
9. Peña-Ayala, A.: Educational data mining: a survey and a data mining-based analysis of recent works. Exp. Syst. Appl. **41**(4), 1432–1462 (2014)
10. Tsai, C.W., Lai, C.F., Chiang, M.C., Yang, L.T.: Data mining for internet of things: a survey. IEEE Commun. Surv. Tutor. **16**(1), 77–97 (2014)
11. Polanco, W.: Web Mining technologies for the e-Commerce solutions in the social networks systems. A Thesis Master of Science - Information Systems, pp. 1–60. SIT, NJ (2013)
12. Milne, D., Witten, I.H.: An open-source toolkit for mining Wikipedia. Artif. Intell. **194**, 222–239 (2013)
13. Beresewicz, M.E.: On representativeness of Internet data sources for real estate market in Poland. Austrian J. Stat. **44**(2), 45–57 (2015)
14. Gök, A., Waterworth, A., Shapira, P.: Use of web mining in studying innovation. Scientometrics **102**(1), 653–671 (2015)
15. Barcaroli, G.: Internet as data source in the Istat survey on ICT in enterprises. Austrian J. Stat. **44**(2), 31–43 (2015)
16. Ferrara, E., De Meo, P., Fiumara, G., Baumgartner, R.: Web data extraction, applications and techniques: a survey. Knowl.-Based Syst. **70**, 301–323 (2014)
17. Kraychev, B., Koychev, I.: Computationally effective algorithm for information extraction and online review mining. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, vol. 64 (2012)
18. Choudhary, S. et al.: Crawling rich internet applications: the state of the art. In: Proceedings of the 2012 Confrence of the Center for Advanced Studies on Collaborative Research, pp. 146–160 (2012)

19. Liu, W., Meng, X., Meng, W.: Vide: a vision-based approach for deep web data extraction. IEEE Trans. Know. Data Eng. **22**(3), 447–460 (2010)
20. Bakaev, M., Avdeenko, T.: Data extraction for decision-support systems application in labour market monitoring and analysis. Int. J. e-Educ. e-Bus. e-Manage. e-Lear. (IJEEEE) **4**(1), 23–27 (2014)
21. Bakaev, M., Avdeenko, T.: Intelligent information system to support decision-making based on unstructured web data. ICIC Expr. Lett. **9**(4), 1017–1023 (2015)