KNOWLEDGE MINING FROM DATA: METHODOLOGICAL PROBLEMS AND DIRECTIONS FOR DEVELOPMENT

JULIUSZ L. KULIKOWSKI

Nałęcz Institute of Biocybernetics and Biomedical Engineering,
Polish Academy of Sciences,
Ks. Trojdena 4, 02-109 Warsaw, Poland
juliusz.kulikowski@ibib.waw.pl

(Received 28 January 2011)

Abstract: The development of knowledge engineering and, within its framework, of *data mining* or *knowledge mining from data* should result in the characteristics or descriptions of objects, events, processes and/or rules governing them, which should satisfy certain quality criteria: credibility, accuracy, verifiability, topicality, mutual logical consistency, usefulness, *etc.* Choosing suitable mathematical models of knowledge mining from data ensures satisfying only some of the above criteria. This paper presents, also in the context of the aims of The Committee on Data for Science and Technology (CODATA), more general aspects of knowledge mining and popularization, which require applying the rules that enable or facilitate controlling the quality of data.

Keywords: data mining, knowledge discovery, data quality, CODATA

1. Introduction

The concept of building an *information society*, which was formulated for the first time in Japan in the 1960s, could have been treated initially as science fiction, however, towards the end of the previous century it became an important element of long-term programmes of social and economic development of the most advanced countries, which was expressed in, among other things, the so-called Bangemann Report [1]. The reason for this was both the globalisation of economic processes and the necessity for taking coordinated international action to protect the natural environment or to prevent natural disasters. Undoubtedly, the interests of large corporations and IT companies constituted (and still constitute) another factor in the development of the information society. A significant body of literature has been devoted to the historical, social and technical aspects of the development of the information society, *e.g.* M. Bazewicz [2], M. Goliński [3],

R. Jacquart (Ed.) [4], J.L. Kulikowski [5], R. Tadeusiewicz [6] and others. The rapid increase of the world's information stores, both in printed and in electronic form, was followed by the development of technical means and organizational rules enabling wide access to and the use of these resources. However, there is a general feeling that these means remain insufficient. The users of information are particularly critical of the *lack of relevance* of the acquired piece of information and of its *redundancy* and of the fact that its *credibility*, *accuracy* and *topicality*, *etc.*, cannot always be guaranteed. The attempts to reduce the redundancy of information by compressing data [7, 8], although allowing to substantially decrease the storage requirements (and at the same time – to economize on material data carriers), do not solve the problem of *semantic redundancy*. Nowadays, the latter problem can be partially solved owing to such techniques as *automatic summarization* of textual information [9], *document indexing* [10], as well as *data mining* [11–14], and *knowledge mining from data* [15–19]. None of these techniques, however, has direct effect on the *quality* of information. Here, the term *quality of information* must be understood in a multifaceted way, since it encompasses a number of qualities, which together affect the usefulness of information for the processes of teaching, researching or decision-making [20, 21]. This paper presents selected problems of ensuring high quality of input data and of the proper application of models for their mining, in order assure the high quality of the knowledge extracted from them.

2. Main directions in knowledge mining from data

Data mining (DM) is defined as a field of IT that deals with methods and programming means of extracting relevant (from the point of view of a given user) information stored implicitly in sets of data. Three things deserve particular attention in this definition: (1) the acquired piece of information should be relevant from the point of view of a given user (or group of users); (2) this piece of information is not explicit, and therefore it is accessible only by means of certain procedures for the analysis of data stored in a database in their original form; (3) data mining can be concerned with data sets of different kind: formatted (*e.g.* tabulated numerical data), partially formatted (*e.g.* images saved in specified formats) or unformatted (*e.g.* texts, music scores, *etc.*). The term *knowledge discovery in databases* (KDD) was coined in the literature in the early 1990s [22] and is currently understood as organized activities aimed at discovering, in large data sets, certain data structures, which are objectively there, have not been recognized before, are practically useful and can be easily interpreted by the user [15]. The interest in both DM and KDD originated in the context of specific needs related to managing companies and due to the observation that operational records of transactions could constitute an important source of information useful for management. A similar observation can also be made in many other areas, such as the operation of medical, educational and public facilities as well as in many fields of experimental research. Such institutions gather data containing scattered and hidden information, which can be extracted, processed using modern computer techniques and then used in practice. The data is stored on a mass scale and maintained for a long time. From the above definitions it follows that DM can be an autonomous operation, but it can also constitute a stage preceding and supporting KDD. In practice, these terms are often treated as synonyms, which, however, blurs the difference between the range of their respective effects, for instance, between calculating the value of the covariance coefficient of two tabulated variables (the result of DM) and determining under what conditions this covariance is positive and large (the result of KDD).

In practice, DM and KDD can be a part of a wider enterprise, aimed at *e.g.* facilitating the operation of a company, taking a specific action, researching a certain phenomenon, *etc.* Therefore, a number of methodologies for designing DM and KDD systems for specific needs [18] are described in the literature. A typical example could be the methodology (proposed by a work team representing several companies) termed *Cross Industry Standard Process for Data Mining* (CRISP-DM) [23]. This methodology distinguishes 6 stages of creating a project, as shown in Figure 1.

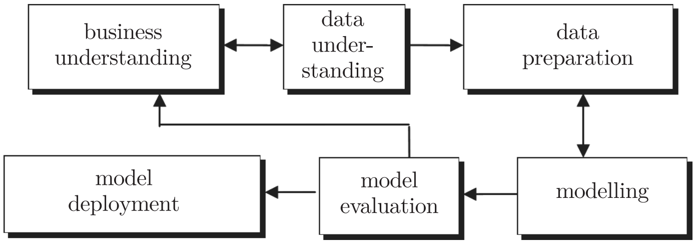

Figure 1. CRISP-DM methodology for designing DM and KDD systems

In this approach, *Business understanding* means the formulation of user's aims and of the requirements related to the DM and KDD system under design. *Data understanding* corresponds to gathering and the initial evaluation of the usability of the data for mining. *Data preparation* consists in formatting the data as required by the mining algorithm. *Modelling* consists in selecting a suitable algorithm for detecting data structures and evaluating their parameters. *Model evaluation* consists in the proper application of the selected algorithms to identify the structures and to evaluate their parameters, as well as in evaluating the conformity of the results with the initial assumptions. Finally, *Model deployment* consists in the complete implementation of the model, including making it available to other users. The schema allows for feedback, and thus for iterativeness. The above-mentioned design stages can also be found (although under different names) in other methodologies for designing DM and KDD systems, and therefore CRISP-DM is considered a *de facto standard* for other, similar solutions [18]. Modified versions of this methodology were also drawn up (CRISP-DM 0.2, CRISP-DM 2007). The procedures described in the methodologies listed here, prescribe the order of activities, but not how they should be carried out, *e.g.*, how the hidden data structures are to be identified. The latter can be realized in practice with formal tools such as:

(1) classical set and relation theory;
(2) classical propositional calculus;
(3) analytic geometry;
(4) harmonic analysis;
(5) approximation theory;
(6) mathematical statistics;
(7) fuzzy and rough set theory;
(8) non-classical logics, *etc.*

The data structures identified using these tools can have the following form:

(1) selected subsets of elements (such as reference objects);
(2) binary or multi-argument relations (similarity, ordering, classification tree structures, *etc.*), hyper-relations;
(3) logical implications;
(4) geometric models;
(5) functional dependencies;
(6) harmonic spectra (or other, functional spectra);
(7) histograms and their parameters (averages, variances, moments of higher order, covariance matrices, *etc.*), as well as probability distributions that approximate them, regression functions, *etc.*;
(8) membership functions of fuzzy sets, fuzzy relations, *etc.*;
(9) statements formulated in the framework of non-classical logics (multivalued, modal, relative, temporal, *etc.*).

3. Ensuring the required data quality

Each of the above DM tools poses specific formal requirements for the input data extracted from databases. In the methodology of CRISP-DM, the *Data understanding* module takes into account the necessity to verify the data quality; however, it does not specify how to achieve this. Therefore, we must differentiate between gathering data for the private purposes of institutions and for the purposes of making them more widely accessible. In the former case, the given institution can control how data is gathered for its purposes and must bear the consequences of potential negligence in this respect. In the latter case, the user has no control over the how the data is gathered and can at best evaluate the usefulness of the available data for their purposes, whereas the responsibility of the organization gathering the data and making it public for the consequences of low data quality is not always clearly specified. When it comes to databases storing data for science, technology and economy, this problem is subject to the particular attention of The Committee on Data for Science and Technology (CODATA), which has its counterparts in over 20 countries (including Poland) and acts under the patronage of the International Council for Science (ICSU). The aim of these organizations is, among others, to create and popularize effective methods of ensuring high quality of the data gathered and made available for the purposes of science and technology. Such data which research, medical, and civil service institutions, *etc.*, gather locally in order to make them available to the public (*e.g.* through Internet access), do not necessarily meet the requirements of their potential users. In general, the following can take place:

(a) the data in the database meet the requirements of semantic correspondence, accuracy and format, determined by the model for exploring the data selected by the user;

(b) the data in the database meet the requirements of semantic correspondence and accuracy, however, their format does not comply with the requirements of the model;

(c) the data in the database do not meet the requirements of semantic correspondence or accuracy required by the model used for data analysis.

Case (a) allows to apply the given formal model to the analysis of the data. Case (b) allows this only after data are suitably reformatted. Case (c) precludes the possibility of data analysis using the given model, since the DM procedures do not allow to reconstruct the information which does not appear in the data. Forcible application of such procedures to data that do not meet the model's requirements of semantic correspondence and accuracy is a serious methodological error, and the results of such data mining should not be deemed credible. However, even case (a) does not guarantee the results of DM to be credible if the source data themselves are not credible, *e.g.* due to methodological errors committed before the data was entered into the database. Such errors can result from the insufficient representativeness of the gathered experimental materials, errors in the utilized measuring technique, the inaccuracy of the system for registering and storing data, *etc.* To protect ourselves from such errors and their consequences, we must register not only the data, but also their meta-information, which allows us to verify them at the source. This can be achieved by storing metadata detailing who, where and when registered the source data, what selection criteria were used when choosing the experimental materials, what measuring equipment or method was used, what methods for the evaluation of the credibility of the measured data were used at the source (*e.g.* rejecting outliers), *etc.* In many cases, however, such metadata are absent from the data made available by the databases. The effect of low quality of the input data on the result of DM and KDD varies between models. It is most visible in the case of deterministic models such as models (1)−(4). Models (5)−(8), by design, concern the analysis of uncertain, incomplete data or data known to contain statistical errors, and their results are in principle given together with error estimates. All DM models are sensitive to hidden systematic errors resulting from unsuitable evaluation methods, improperly calibrated measuring equipment, *etc.* The only and not necessarily available method of detecting (but not correcting) such errors is to assess their consistency by comparison with other data or with acceptable ranges of the data values; this also applies to the possibility of the user detecting out-of-date data (if it has not been detected via the use of metadata).

The proper interpretation of the assumptions underlying the applied model is also crucial for the quality of DM and KDD. Common methodological errors include:

- uncritical application of statistical models of regression, estimation or verification of hypotheses suitable for normally distributed random variables to variables that are evidently not normally distributed (for instance, exhibiting strong asymmetry);
- uncritical application of the methods of object classification based on distance metrics (*e.g.* Euclidean, Manhattan, Chebyshev distance) without regard for dimensional analysis and neglecting the fact that quantities can be expressed in different units (which has an effect on the relative impact of different constituents on the classification result);
- improper interpretation of certain models (*e.g.* interpreting a regression curve as a cause and effect relationship of a pair of variables, when in fact they are merely dependent on a third variable).

4. Conclusions

Data mining and knowledge mining from data (or knowledge discovery in databases) play an increasingly more important role in recovering valuable information from large sets of data. However, the effective use of such tools requires observing certain rules and properly selecting and controlling the quality of input data. This, in turn, necessitates the co-operation of the entities that make the data available to users.

References

[1] Bangemann M 1994 *Europe and the Global Information Society. Recommendations to the European Council*, Bruksela (in Polish)

[2] Bazewicz M 2000 *A Vision of Communication, Information and Knowledge Society in XXI Century*, SILESIA, Wrocław (in Polish)

[3] Golinski M 1997 *Development Level of Information Infrastructure of a Society. An Attempt to Evaluation*, AOW PLJ, Warsaw (in Polish)

[4] Jacquart R (Ed.) 2004 *Building the Information Society. IFIP 18th World Computer Congress, 22–27 August 2004*, Kluwer Academic Publishers, Boston

[5] Kulikowski J L 1978 *Information and the World Where We Live*, WP, Warsaw (in Polish)

[6] Tadeusiewicz R 2002 *The Society of Internet*, AOW EXIT, Warsaw (in Polish)

[7] Skarbek W (Ed.) 1998 *Multimedia. Algorithms and Compression Standards*, AOW PLJ, Warsaw (in Polish)

[8] Bhaskaran V and Konstantinides K 1995 *Image and Video Compression Standards. Algorithms and Applications*, Kluwer Academic Publishers, Boston

[9] Kacprzyk J, Yager R R and Zadrożny S 2000 *Int. J. of Appl. Math. and Comp. Sci.* **10** 813

[10] Ouziri M, Verdier C and Flory A 2003 *Intelligent Information Processing and Web Mining*, (Kłopotek M A, Wierzchoń S T and Trojanowski K, Eds), AiSC, Springer-Verlag, Berlin, pp. 189 – 198

[11] Michalski R S 1994 *Seeking Knowledge in the Flood of Facts. Intelligent Information Systems*, Proc. of the Workshop held in Wigry, Poland, 6–10 June, 1994, IPI PAN, Warsaw

[12] Weiss S M and Indurkhya N 1998 *Predictive Data Mining. A Practical Guide*, Morgan Kaufmann Publishers, Inc., San Francisco

[13] Cios K J (Ed.) 2001 *Medical Data Mining and Knowledge Discovery. Studies in Fuzziness and Soft Computing*, Physica-Verlag, Heidelberg

[14] Coenen F 2011 *The Knowledge Eng. Rev.* **26** (1) 25

[15] Maimon O and Rokach L 2005 *The Data Mining and Knowledge Discovery Handbook*, Springer, New York

[16] Larose D T 2006 *Discovering Knowledge in Data. An Introduction to DATA MINING*, WN PWN, Warsaw (in Polish)

[17] Jashapara A 2006 *Knowledge Management: an Integrated Approach*, PWE, Warsaw (in Polish)

[18] Mariscal G, Marban Ó and Fernandez C 2010 *The Knowledge Eng. Rev.* **25** (2) 137

[19] Chen H, Fuller S S, Friedman C and Hersh W 2005 *Medical Informatics. Knowledge Management and Data Mining in Biomedicine*, IS 2, Springer, USA

[20] Kulikowski J L 2009 *Data Quality Assessment. Innovations in Database Technologies and Applications. Current and Future Trends*, (Ferragine V E, Doorn J H and Rivero L C, Eds) Vol. I, Chapt. XLI, Information Science Reference, Hershey

[21] Shankaranarayanan G and Even A 2009 *Measuring Data Quality in Context. Innovations in Database Technologies and Applications. Current and Future Trends*, (Ferragine V E, Doorn J H and Rivero L C, Eds) Vol. I, Chapt. XLII, Information Science Reference, Hershey

[22] Piatetsky-Shapiro G 1991 *Report on the AAAI-91 Workshop on Knowledge Discovery in Databases*, Technical Report 6, IEEE Expert

[23] Chapman P *et al.* 2000 *CRISP-DM 1.0 Step-by-Step Data Mining Guide*, Technical Report, CRISP-DM