Towards a Meaningful Analysis of Big Data
Enhancing Data Mining Techniques through a Collaborative Decision Making Environment

Nikos Karacapilidis[1], Stefan Rueping[2], Georgia Tsiliki[3] and Manolis Tzagarakis[1]
[1]*Computer Technology Institute and Press "Diophantus", University of Patras, 26504 Rio Patras, Greece*
[2]*Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany*
[3]*Biomedical Research Foundation, Academy of Athens, 115 27 Athens, Greece*

Keywords: Big Data, Data Mining, Decision Support Systems, Modeling and Managing Large Data Systems.

Abstract: Arguing that dealing with data-intensive settings is not a technical problem alone, we propose a hybrid approach that builds on the synergy between machine and human intelligence to facilitate the underlying sense-making and decision making processes. The proposed approach, which can be viewed as an innovative workbench incorporating and orchestrating a set of interoperable services, is illustrated through a real case concerning collaborative subgroup discovery in microarray data. Evaluation results, validating the potential of our approach, are also included.

1 INTRODUCTION

Generally speaking, the Web 2.0 era is associated with huge, ever-increasing amounts of multiple types of data, obtained from diverse and distributed sources, which often have a low signal-to-noise ratio for addressing the problem at hand. Nowadays, it is easier to get the data in than out. Big volumes of data can be effortlessly added to a database (e.g. in transaction processing); the problems start when we want to consider and exploit the accumulated data, which may have been collected over a few weeks or months, and meaningfully analyze them towards making a decision. Admittedly, when things get complex, we need to identify, understand and exploit data patterns; we need to aggregate big volumes of data from multiple sources, and then mine it for insights that would never emerge from manual inspection or analysis of any single data source. In other words, the pathologies of big data are primarily those of analysis. In contemporary information management settings, the way that data are structured for query and analysis, as well as the way that tools are designed to handle them efficiently, are crucial issues that certainly set a big research challenge.

Focusing on the meaningful analysis of big data, this paper presents an innovative approach being implemented in the context of the EU funded Dicode research project (http://dicode-project.eu/). Arguing that dealing with data-intensive and cognitively complex settings is not a technical problem alone, Dicode adopts a hybrid approach that builds on the synergy between machine and human intelligence to facilitate the underlying sense-making and decision making processes.

Overall, the proposed approach aims to exploit the information growth by (i) ensuring a flexible, adaptable and scalable information and computation infrastructure, coupled with the appropriate data mining techniques, and (ii) exploiting the competences of stakeholders and information workers to meaningfully confront information management issues such as information characterization, classification and interpretation, thus enhancing - in various ways - the utilization of data mining techniques. Exploitation of stakeholders' competences is vital in contemporary decision making settings. As the amount of available resources increases and becomes more specialized, stakeholders tend to form teams and seek collaboration with their peers to address complex questions. In other words, such settings become increasingly interdisciplinary and collaborative in nature (Hara et al., 2003); (Schur et al., 1998).

The remainder of this paper is structured as follows: Section 2 reports briefly on relevant

Karacapilidis N., Rueping S., Tsiliki G. and Tzagarakis M..
Towards a Meaningful Analysis of Big Data - Enhancing Data Mining Techniques through a Collaborative Decision Making Environment .
DOI: 10.5220/0004051201410146
In *Proceedings of the International Conference on Data Technologies and Applications* (DATA-2012), pages 141-146
ISBN: 978-989-8565-18-1

requirements and challenges in the fields of data-mining and collaboration support. Section 3 presents the overall approach adopted in the Dicode project towards addressing these problems. The proposed approach is illustrated in Section 4, through a specific example concerning interactive subgroup discovery in microarray data. Evaluation results for the related Dicode services are sketched in Section 5. Finally, concluding remarks and future work directions are outlined in Section 6.

2 CHALLENGES AND REQUIREMENTS

Generally speaking, information management related tasks need to be streamlined and automated. Recent findings clearly indicate that information management costs too much when it is not well organized and meaningfully automated (Eppler and Mengis, 2004). They also call for investments in innovative software that reduces or eliminates time wasted, reduces management overheads, stream-lines collaborative processes, and automates the overall workflow. Return on such investments can be both tangible (e.g. time or money saved) and intangible (e.g. more valuable information, easier extraction of hidden information, increase of information workers' satisfaction and creativity, improved collaboration).

As results from the above, issues related to the guidance of the information worker through the space of available data and the proper interpretation of relevant information to augment the corresponding decision making activities are of major importance. Towards this direction, we foresee a semi-automatic, adaptive approach that makes use of semantic metadata and pre-structured data patterns to provide plausible recommendations, while also learning from the users' feedback to better target their information interests (Adomavicius and Tuzhilin, 2005). This will be enabled by innovative data mining techniques and services such as local pattern mining, similarity learning, and graph mining, together with a flexible framework where all these services are seamlessly integrated and orchestrated. Current research in data-mining and collaborative decision making in data intensive settings impose new sets of requirements. In data-mining, for example, since data and information is today available in large volumes and diverse types of representation, intelligent integration of these data sources to generate new knowledge (towards serving collaboration and decision making requirements) remains a key challenge. Moreover, contemporary approaches need to help users utilizing complex multi-source data in a reasonable way by supporting them in finding relevant information and by providing personalized recommendations. Another big category of requirements concerns the exploration, delivery and visualization of the pertinent information. Related to collaborative decision making support, stakeholders require innovative services that shift in focus from the mere collection and representation of large-scale information to its meaningful assessment, aggregation and utilization in contemporary collaboration and decision making settings.

3 THE DICODE APPROACH

Dicode provides a Web-based platform with advanced data mining and collaborative decision making support services. Central to the Dicode approach is the concept of "workbench". Workbenches enable the seamless integration of various data mining and collaborative decision making support services (Figure 1). They can be configured by end users as far as the services to be included are concerned (according to the needs of the particular context and problem under consideration). A widget-based approach has been adopted to offer the relevant services. Services provided through the Dicode platform are:

- *Data acquisition services*, which enable the purposeful capturing of tractable information that exists in diverse data sources and formats on the web.
- *Data pre-processing services*, which efficiently manipulate and transform raw data before their storage to the foreseen solution.
- *Data mining services*, which exploit and are built on top of a cloud infrastructure and other most prominent large data processing technologies to offer functionalities such as high performance full text search, data indexing, classification and clustering, directed data filtering and fusion, and meaningful data aggregation. Advanced text mining techniques such as named entity recognition, relation extraction and opinion mining aid the extraction of valuable semantic information from unstructured texts. Intelligent data mining techniques being elaborated include local pattern mining, similarity learning, and graph mining.
- *Collaboration support services,* which facilitate

the synchronous and asynchronous collaboration of stakeholders through adaptive workspaces, efficiently handle the representation and visualization of the outcomes of the data mining services (through alternative and dedicated data visualization schemas).

- *Decision making support services*, which augment both individual and group sense- and decision-making by supporting stakeholders in locating, retrieving and arguing about relevant information and knowledge, as well as by providing them with appropriate notifications and recommendations (taking into account parameters such as preferences, competences, expertise etc.). Services belonging to this category primarily exploit the reasoning capabilities of humans.

4 AN EXAMPLE: INTERACTIVE SUBGROUP DISCOVERY IN MICROARRAY DATA

Focusing on a particular data mining technique (namely, Subgroup Discovery), and through a real use case concerning collaborative work with microarray data, this section reports on the utilization and interoperation of Dicode services.

4.1 Subgroup Discovery

Subgroup Discovery (Kloesgen, 1996) is the task of finding patterns that describe subsets of a database with a high statistical unusualness in the distribution of a target attribute. For example, in a group of patients that did or did not respond to specific treatment, an interesting subgroup may be that patients who are older than 60 years and do not suffer from high blood pressure respond much better to the treatment than the average. Subgroup Discovery is a popular approach for identifying interesting patterns in data, because it combines statistical significance with an understandable representation of patterns.

Algorithmically, subgroup discovery ranks possible subgroups according to a quality function q, which depends on both the size of the subgroup (in the example above, the number of patients in the study over 60 years without high blood pressure), and the probability of the target attribute in the subgroup (the percentage of patients in the subgroup that responded well to the treatment). That is, subgroup discovery prefers subgroups where the distribution of the target attribute is unusually high - such that the results are interesting - but where at the same time the subgroup is large, such that the results are reliable. A popular choice for the quality function is to use the statistical significance of the subgroup as calculated by a binomial test. Given a set of descriptive attributes (e.g. age and diagnosis), the space of possible subgroups consists of all rules that can be formed by conjunctions of attribute-values comparisons. Subgroup Discovery tests all possible subgroups and reports the top k subgroups with respect to the given quality function. Efficient algorithms for fast subgroup discovery exist (Grosskreutz et al., 2008).

Since Subgroup Discovery is often used to generate a human-understandable representation of the most interesting dependencies in a data set, the practical usefulness of the approach depends on the ability to produce a concise and interesting output. Hence, in the practical application of this approach, care must be taken to select the proper descriptive attributes. Frequent mistakes that are made in practice are that informative attributes are left out (such that no good subgroup can be discovered), or that undesired attributes are included. For example, undesired attributes in the context of clinical decision making may be attributes that are only known after the treatment of a patient and not at the point of time where the decision about the treatment must be taken.

Subgroup Discovery is especially well suited for tasks where there exist concise descriptions of the patterns in the data, but where it is not easy to find these descriptions due to the high number of attributes and size of the data. Hence, to apply Subgroup Discovery to gene data, it is important to generate enriched, highly informative attributes. For example, Trajkoski and his colleagues (Trajkovski et al., 2006) describe an approach that enriches microarray data with descriptions from the Gene Ontology, which serves to summarize the commonalities of the many genes that are found in a traditional study.

It is obvious that the inclusion of more features or the identification of undesired ones (a task that may not always be simple due to the many correlations in real-world patient data) is a process which requires much background knowledge by the involved biologists and clinicians. Hence, when one thinks about clinical knowledge discovery based on data mining tools such as Subgroup Discovery, it is important to set up a collaborative, interactive process, where users decide about which data repositories should be considered, analyze the algorithmic results, discuss the weaknesses of the

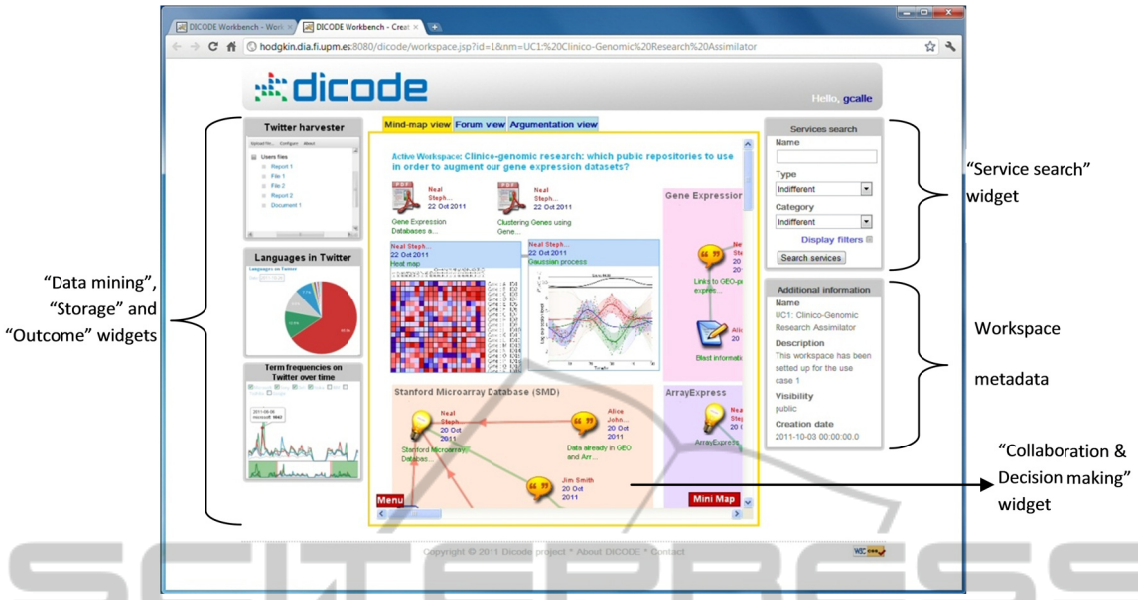

Figure 1: Screenshot of the Dicode workbench. The central widget provides access to collaboration support services. On the right, the search service widget allows users to locate new services. Widgets on the left are customizable by the user and concern data mining services.

patterns that were identified, and set up a new iteration of the algorithm by defining other descriptive attributes or integrating other relevant data. Technically, this can be facilitated through interaction between the users and the Subgroup Discovery algorithm in various ways, such as:

- presenting the discovery patterns to the users in a form that allows their use as one piece of knowledge in the overall discussion process;
- making it easy for the users to give feedback to the algorithm, e.g. by allowing them to specify undesired attributes, non-interesting subgroups, or controlling the complexity of the output, such as the desirable length of the rules (Rüping, 2009);
- making it easy in the overall system to select (and integrate new) data sets and attributes, both from external data sources as well as from the underlying discussion.

Subgroup Discovery is one of the data mining services offered through a user-friendly widget in the Dicode approach. Through appropriate functionalities, this widget may interoperate with the widget offering the collaborative decision making service (middle part of Figure 1). This last service is described in the next subsection through an example scenario and can achieve the desired interaction mentioned above. The example illustrated concerns collaboration towards deciding the most appropriate data repository to be considered in the Subgroup Discovery technique.

4.2 Collaborative Decision Making

Two researchers, Jim (bioinformatician) and Alice (biologist), aim to investigate which genes or groups of genes are associated with breast cancer (BC) disease. Jim and Alice have independently conducted similar analysis with in-house gene-expression datasets; however, their findings were not very encouraging, which was attributed to the small sample size (i.e. number of patients) available. They decide to search the available public repositories and download suitable data to augment their data in order to produce more reliable results. Niall (data manager) suggests the following alternatives (in alphabetic order): (i) Array Express (ArEx), (ii) Gene Expression Omnibus Datasets (GEO), and (iii) Stanford Microarray Database (SMD).

All three of them are good candidates being public repositories that archive and freely distribute high-throughput multi-platform gene-expression data, submitted by and following the standards of the scientific community. Jim and Alice are aware of all three databases and start discussing each available alternative by commenting in favour or against each option by indicating for example whether they are highly populated, provide supplementary files such as platform information and clinical data and they allow packages to retrieve the data and include tools for first analysis of the data. Jim and Alice may also comment on the issues raised. Based on the course of the discussion, they can discard alternative

solutions and finally decide collectively about which repository to use.

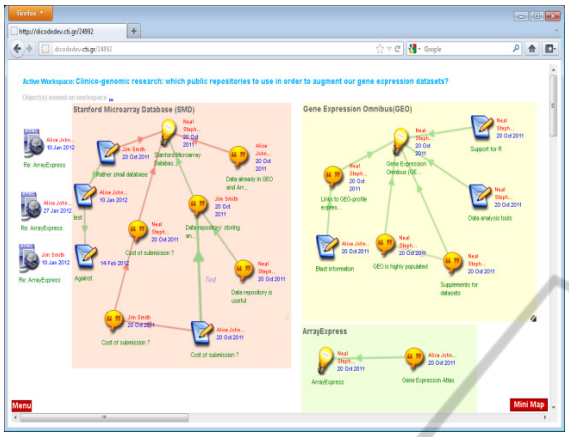

Figure 2: An instance of the Dicode's Collaborative Decision Making Support Service (mind map view) for the case reported in Section 4.2.

4.3 Collaborative Decision Making Support in Dicode

Figure 2 shows an instance of the Dicode's Collaborative Decision Making Support Service for the collaboration reported in the previous subsection. In this instance, the collaboration space is displayed in a "mind-map view", where users can interact with the items uploaded. In this view, stakeholders (Jim, Alice and Niall) may organize their collaboration through dedicated item types such as *ideas*, *notes* and *comments*. Ideas stand for items that deserve further exploitation; they may correspond to an alternative solution to the issue under consideration and they usually trigger the evolution of the collaboration. Notes are generally considered as items expressing one's knowledge about the overall issue, an already asserted idea or note. Finally, comments are items that usually express less strong statements and are uploaded to express some explanatory text or point to some potentially useful information. Multimedia resources can also be uploaded into the mind map view (the content of which can be directly embedded in the workspace). The mind-map view deploys a spatial metaphor permitting the easy movement and arrangement of items on the collaboration space. All available items can be interrelated by directed arrows. Coloured rectangles are visual conveniences to group similar items and enable the implementation of appropriate abstraction mechanisms (Shipman and McCall, 1994). Using the mind-map view, stakeholders can transform the resources from a mere collection of items into coherent knowledge structures that facilitating sense making on the available resources.

By using the search facilities of the workbench, they are also able to search for relevant literature or data sets, which can be also uploaded on the collaboration logbook (by drag-and-drop). Moreover, resources that stakeholders agree that are relevant for their research can be added to the sources section of the workbench (by dragging and dropping them from the logbook into the respective section of the workbench).

Stakeholders may need to further elaborate the knowledge items considered so far, and exploit additional functionalities to advance their argumentative collaboration towards reaching a decision. Such functionalities can be provided by the "formal view" that – based on the IBIS discourse model (Kunz and Rittel, 1970) – enables the semantic annotation of knowledge items, the formal exploitation of collaboration items patterns, and the deployment of appropriate formal argumentation and reasoning mechanisms which determines the status of each discourse entry, the ultimate aim being to keep stakeholders aware of the discourse outcome. A detailed description of the formal view can be found in (Karacapilidis and Papadias, 2001). While a mind map view aids the exploitation of information by stakeholders (i.e. makes the logbook mainly human-interpretable), a formal view aims mainly at the exploitation of information by the machine (i.e. makes the logbook mainly machine-interpretable). By switching from mind map into formal view, existing item types are transformed, filtered out, or kept "as-is" based on a specific set of rules.

5 EVALUATION

Dicode services go through an ongoing evaluation process. The first evaluation round of Dicode services mainly aimed to assess a series of key success indicators concerning the maturity of the technology used, as well as the usability and acceptability of Dicode services in three real-life contexts (clinic-genomic research, medical decision making, and opinion mining from Web 2.0 data). Evaluators were asked to read a service-specific scenario, experiment with the Dicode services and fill in a mixed-type questionnaire (responses expected were in both quantitative and qualitative form). For the case reported in this paper, the sample consisted of 58 evaluators with varying background knowledge in bioscience fields. Answers to the quantitative questions of the questionnaires were

given in a 1-5 scale, where 1 stands for 'I strongly disagree' and 5 for 'I strongly agree'.

With respect to the overall quality of the Dicode Workbench, the evaluators agreed (median: 4, mode: 4) that its objectives are met, that it is novel to their knowledge, that are satisfied with its performance and that they are overall satisfied with it. The evaluators were neutral (median: 3, mode: 3) with respect to whether the Workbench addressed the data intensive decision making issues. As long as its acceptability is concerned, the evaluators agreed (median: 4, mode: 4) that the Workbench has the full set of functions they expected, that its interface is pleasant and that they will recommend it to their peers/community.

6 CONCLUSIONS

Taking into account the feedback received from the first evaluation phase of the Dicode project, we argue that our overall approach offers an innovative solution that reduces the data-intensiveness and overall complexity of real-life collaboration and decision making settings to a manageable level, thus permitting stakeholders to be more productive and concentrate on creative activities. Towards this direction, the project provides a suite of innovative, adaptive and interoperable services that satisfies the requirements reported in Section 2.

A major future work direction concerns the improvement of Dicode services in terms of their documentation, user interfaces and performance. Another one concerns testing of these services in various data-intensive contexts towards further assessing their applicability and potential.

ACKNOWLEDGEMENTS

This publication has been produced in the context of the EU Collaborative Project "DICODE - Mastering Data-Intensive Collaboration and Decision Making", which is co-funded by the European Commission under the contract FP7-ICT-257184. This publication reflects only the authors' views and the Community is not liable for any use that may be made of the information contained therein.

REFERENCES

Adomavicius, G., Tuzhilin, A., 2005. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering* 17(6), pp. 734–749.

Eppler, M. J., Mengis, J., 2004. The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines. *The Information Society* 20, pp. 325–344.

Grosskreutz, H., Rüping, S., Wrobel, S., 2008. Tight Optimistic Estimates for Fast Subgroup Discovery. In: *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Springer LNAI.

Hara, N., Solomon, P., Kim, S., Sonnenwald, H., 2003. An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information Science and Technology* 54(10), pp. 952–965.

Karacapilidis, N., Papadias, D., 2001. Computer Supported Argumentation and Collaborative Decision Making: The HERMES system. *Information Systems* 26(4), pp. 259-277.

Kloesgen, W., 1996. Explora: A Multi-pattern and Multi-strategy Discovery Assistant. In: *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Kunz, W., Rittel, H., 1970. Issues as Elements of Information Systems. Technical Report 0131, Universität Stuttgart, Institut für Grundlagen der Planning.

Rüping, S., 2009. Ranking Interesting Subgroups. In: L. Bottou and M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*. Omnipress, pp. 913-920.

Schur, A., Keating, K. A., Payne, D. A., Valdez, T., Yates, K. R., Myers, J. D., 1998. Collaborative suites for experiment-oriented scientific research. *Interactions* 3(5), pp. 40–47.

Shipman, F. M., McCall, R., 1994. Supporting knowledge-base evolution with incremental formalization. In: *Proceedings of the CHI '94 Conference*. pp. 285-291.

Trajkovski, J., Zelezny, F., Tolar, J., Lavrac, N., 2006. Relational Subgroup Discovery for Descriptive Analysis of Microarray Data. In: M. Berthold, R. Glen, I. Fischer (Eds.), *Computational Life Sciences II*. Springer, pp. 86-96.