Supporting Exploratory Search by User-Centered Interactive Data Mining

Stefan Haun, Andreas Nürnberger
Data & Knowledge Engineering Group, Faculty of Computer Science,
Otto-von-Guericke-University Magdeburg, Germany
{stefan.haun,andreas.nuernberger}@ovgu.de

ABSTRACT

A core task in e-discovery scenarios is to explore and mine large distributed heterogeneous information archives. Therefore, very often an exploration and analysis of on-the-fly generated subsets, gathered e.g. from mining, filtering or search operations, is necessary. This requires performant and user-friendly ways to interactively integrate, mine and explore. However, Data Mining and Human Computer Interaction seem to be two worlds of their own: Batch processing and complex, i.e. long-running algorithms on one side, interactivity and quick response times on the other side. Exploratory tools for e-discovery approaches make it necessary to seamlessly combine data mining algorithms and user interaction. Therefore, we identify and discuss some open questions towards closing this gap and making data exploration and mining more user-friendly.

1. INTRODUCTION

E-discovery tools support a user in finding information in large data sets, such as personal information, e.g. e-mail, large document sets or the World Wide Web. These tools use methods from both, Data Mining and Human-Computer Interaction, which themselves have very different requirements. We identify and discuss some open questions which should be considered when designing and building usable e-discovery solution: Interactivity, exploratory search using data mining methods and entity identification so support tool and data integration. Finally we introduce a tool that supports graph exploration and loose coupling of data sources and mining algorithms.

2. RELATED WORK

Related work towards graphical user interfaces for data mining can be found in tools like the *Information Miner*[8], *KNIME*[2] or *WEKA*[7], which provide a integrated environment for creating and running data mining workflows. *Polaris*[10] and *Jigsaw*[4] are designed for Visual Analytics and require less expert knowledge in the data mining methodologies, the latter with a strong emphasis on e-discovery.

A description of search user interfaces can be found in [6], while the basic principles of user interface design—as described by Shneiderman[9]—should also be followed by e-discovery tools.

[1] and [3] elaborate the question of combining Data Mining or E-Discovery and Human-Computer Interaction.

3. OPEN ISSUES

Data Mining and User Interfaces are two worlds of their own. While Data Mining processes are often complex and long-running, User Interfaces need to be fast and responsive. As with E-Discovery, these two worlds begin to merge and we need a bridge between opposite environments. Starting points can be the following requirements:

3.1 Making Data Mining more interactive

Data Mining often uses batch processing, i.e. a script or workflow is run until it finished with some result. There are user interfaces for data mining, but they are mostly meant to support building workflows and execute them as batch processing. However, the size and content of the result set cannot be forseen and processing time may vary between fractions of a second and days.

User Interfaces, on the other hand, must have quick and reliable response times (with a maximum between 2–4 seconds). The user must be able to anticipate the behaviour of the function just called and its outcome to support planning for the task at hand. For long processes, the progress must be shown and often an estimation of the time to finish the task is required. Also, the user may want to cancel a progress or change parameters in between to adapt to changes in the environment or her needs for the outcome.

As a consequence, data mining processes need to become more interactive.

In order to allow a user to decide whether she wants to cancel a process or change its parameters intermediary results are necessary. However, data mining algorithms are often monolithic and those results are hard to acquire. Those algorithms need to be broken up to have more "entry points" which allow to inspect the so-far results and, on occasion, change the parameters. Iterative algorithms with monotonic error functions can support this very well: instead of simply running the algorithm until a minimum error is reached, results may be returned after a set of iterations so that the user can decide whether she wants the calculation to be continued or whether the results are accurate enough. When intermediary results are not possible or the time taken cannot be estimated, statistical runtime obervation may help to give the user a clue about mean running times for this kind of task. It may still be better to tell whether a task takes two minutes rather than half an hour instead of leaving the user absolutely clueless.

SIGIR'11, July 24–28, 2011, Beijing, China.
ACM X-XXXXX-XX-X/XX/XX.

3.2 Supporting Exploratory Search using Data Mining Methods

Data items can be frequently visualized as connected information entities in form of graphs. Even if links are not explicitly available, they can very often be obtained, e.g. by using similarity measures or by exploiting meta-data to connect very similar entities. Once the graph stucture is obtained—or on-the-fly generated during user interaction—the graph itself can be used to guide a user in his exploration process. As an advantage the user does not need to know exactly what she is looking for, but instead may begin with a promising starting point and refine the search or mining parameters according to intermediary results. However, exploration incorporates much interaction with very short loops, i.e. it requires data mining methods which have a very quick response time, but only need to return small result sets. Indexes can provide a quick access using precalculated data. However, these calculations cannot take the user's context into account. If a data mining method can be broken into global—or invariant—and context-dependent parts, those calculations could be carried out much faster.

3.3 Entity Identification in Data Integration

If an integrated view on data mining results is provided, sub-graphs or sub-sets can be used to subsequently build and manipulate a result set from different sources and with different tools. A problem appears with the entity identification: There must be a consensus between all tools about how to identify equal entities. This can be achieved by finding a global convention for naming entities which takes different contexts and knowledge domains as well as different information representations into account. In the *Semantic Web* this problem is often solved with *Uniform Resource Identifiers* (URIs). However, there is no perfect solution towards how these URIs should be built. For resources available on the WWW, the *Uniform Resource Locator* (URL) can be used, if it is in both ways unambigous. For other entities it is hard to achieve a suitable standard. Either there ie no conceiveable URI—e.g. what should be the URI of a specific appointment?—or there are multiple solutions, e.g. when persons are identified with their e-mail address, a web page or their phone number. If possible, there should be a common solution acceptable for all users. Otherwise, a workaround can be mapping agents, i.e. tables or algorithms that map between entity identifications to find similarities, e.g. to find overlapping nodes while merging sub-graphs.

4. EXPLORATORY SEARCH AS GRAPH NAVIGATION

One possible approach to better integrate data mining and explorative search processes is to loosely couple both in an interactive graph browser. A first step in this direction is the *Creative Exploration Toolkit (CET)* is a user interface with several distinct features: Support of interactive graph visualization and exploration, integration of external data providers for arbitrary linked data sources integration of a modular open source data analytics system and easy configuration to serve specific user requirements. A brief description of the tool can be found in [5].

Instead of loading a complete graph for visualization, the shown graph can be built accumulatively, i.e. the graph in the user interface is created by merging several subsequently received sub-graphs, resulting in a graph which has evolved from the user's interactions with one or more datasets and algorithms.

The above issues have been discovered during the development of the CET. While it is relatively easy to integrate different tools and get results from different data sources, it is still quite hard to integrate these results and keep the user appraised about what is going on behind the scene.

5. CONCLUSION

In order to provide good, usable e-discovery tools, we need user interfaces which can convey progress and status of data mining processes and data mining processes which are interactive, i.e. which can be cancelled or have their parameters changed and are less monolithic. We have identified three aspects we think are important towards "user-friendly data mining" and shortly introduced a tool—the CET—which is our platform for researching these issues.

We would like to start a discussion about good solutions for the mentioned issues to improve the usability of E-Discovery tools.

6. REFERENCES

[1] S. Attfield, S. De Gabrielle, and A. Blandford. The Loneliness of the Long-Distance Document Reviewer: E-discovery and Cognitive Ergonomics. *DESI III Global E-Discover/E-Disclosure Workshop: 12th International Conference on Artificial Intelligence and Law*, 2009.

[2] M. R. Berthold, N. Cebron, F. Dill, G. D. Fatta, T. R. Gabriel, F. Georg, T. Meinl, P. Ohl, C. Sieb, and B. Wiswedel. Knime: The Konstanz Information Miner. Technical report, 2006.

[3] J. G. Conrad. E-discovery revisited: A broader perspective for ir researchers, 2007.

[4] C. Görg and J. Stasko. Jigsaw: investigative analysis on text document collections through visualization. In *DESI II: Second International Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings.* University College London, UK, 2008.

[5] S. Haun, A. Nürnberger, T. Kötter, K. Thiel, and M. Berthold. CET: a tool for creative exploration of graphs. *Machine Learning and Knowledge Discovery in Databases*, pages 587–590, 2010.

[6] M. A. Hearst. *Search User Interfaces.* Cambridge University Press, 1 edition, 2009.

[7] G. Holmes, A. Donkin, and I. Witten. WEKA: A Machine Learning Workbench. In *Proc Second Australia and New Zealand Conference on Intelligent Information Systems*, Brisbane, Australia, 1994.

[8] F. Rügheimer and R. Kruse. Information Miner – a Data Analysis Platform, 2005.

[9] B. Shneiderman. *Designing the User Interface.* Addison-Wesley, 1998.

[10] C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional databases. *Commun. ACM*, 51:75–84, November 2008.