IST WORLD – MACHINE LEARNING AND DATA MINING AT WORK

Jure Ferlež
Department of Knowledge Technologies and Center for Knwoledge Transfer, Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 477 3127; fax: +386 1 477 3315
e-mail: Jure.Ferlez@ijs.si

ABSTRACT

This paper is an overview of practical machine learning and data mining solutions used in the IST World portal (*http://www.ist-world.org/*). IST World portal is a customized data mining application for mining research related information. Machine learning and data mining methods are used to tackle the problem of data integration and data analysis. This includes automatic classification, unsupervised clustering, multi dimensional scaling, graph centrality measures and graph drawing algorithms.

1 INTRODUCTION

In this paper we overview the data integration and analysis functionality of IST World portal (http://www.ist-world.org/) [1]; which is a web application built on top of several machine learning and data mining algorithms.
The portal integrates information about research and technology development actors such as organizations and experts on a local, national and European level. The functionality of the portal [2] aims at providing the user with assistance on discovering competence and collaboration knowledge on researchers and research organizations. It also shows the context of their co-operation in the joint projects and publications.
To meet the described functionalities the portal is setup in the form of a data mining application specifically customized to discovering knowledge about *research projects*, *research papers*, *research organizations* and *researchers* [3] (In the rest of the paper we shall name all of these as *entities*). The typical scenario of usage of the application consists of two steps: (1) complex search and navigation step to retrieve relevant information from the data repository, (2) automated analytical and visualization step to present the results. So first, in a selection step the entities (organizations, experts, publications, projects) or subsets of the entities that represent the target of interest are to be specified by the user by making use of available search and navigation functionalities. Second, in an analysis step one of the analytical methods is applied upon the retrieved set of entities and the results are presented by a visualization technique.

In order to implement the described knowledge discovery functionality, data mining and machine learning approach proves necessary. We use it (1) to integrate the data imported from various data sources, (2) to perform an off-line time demanding data analysis and (3) to provide the efficient on-line analytic tools. The majority of the used machine learning algorithms can be classified in to the fields of either text mining or graph mining.
Section 2 of this paper describes the data integration process and the machine learning and data mining techniques which are used. Section 3 describes the offline time demanding analysis of the data providing information on the entities in the global context of all the data in the repository. Section 4 describes the on-line analytical tools which are used to discover knowledge on individual entities in the context of a selected set of entities.

2 DATA INTEGRATION

Multiple data sources of the IST World portal cause the problem of data integration. The data in the IST World portal repository is imported from different types of public data sources like national research databases, global research paper databases or other sources like web crawls [4]. Therefore the evident problem is identifying and linking of the records from different data sources that describe the same entity. The following content of this section describes the integration approach we think is best suited for the needs of the portal. This integration approach will be implemented in the near future.
This problem of finding and resolving duplicate records is also known as Record Linkage problem or Object Identification problem [5]. This is a well known problem and many solutions have already been suggested [5]. We plan to use the following standard approach to solve the problem: (1) we will produce a list of the candidate pairs of records for merging. This is a standard step of Record Linkage process also called blocking [5]. There exist many different solutions for it. (2) We will decide on a given candidate pair whether to merge it or not. This step of the Record Linkage process is called matching [5] and many solutions for it exist as well.

In the context of IST World portal we will seek for the solution that best fits in the context of the data to be integrated. The integration problem in IST World portal is

a consequence of different data sources describing the same *entity*, which results in two or more database records about the one *entity*. The two records from the two data sources should therefore be combined into one record of the *entity*. What is important that the task at hand is about merging databases with mostly correctly typed names and therefore the biggest problem at hand is the blocking of the entities. Both of the Record Linkage steps in the context of the IST World portal are supported by the use of machine learning and data mining techniques customized to the context of IST World as described bellow.

2.1 Blocking

It is our belief that since the data to be integrated in the IST World Portal originates from well maintained databases, that the blocking part of the linking process is the heavier part of the problem. Therefore we try to exploit the full text indexing and full text querying of the records stored inside relational database in a novel way to identifying merge candidate pairs quickly and efficiently. The full text querying allows us to search for records that have approximately the same name. Thus we will produce a merge candidate record based on the full text indexing similarity of the names of *entities*. The full text indexing allows us to match records with the same words or same inflection of the words regardless of their order. However it does not allow us to match any records that may have been mistyped and therefore do not share exactly the same words. This technique is thus most suitable for records with the correct spelling. As said, most of the data in the IST World repository originates from well maintained databases. This makes the problem of merging of two entities with correctly spelled names much more frequent than problem of merging of records with spelling errors in their names. This later problem is mostly the problem of merging the data entered into the database by hand. This kind of data however only represents very small part of the data to be integrated and is therefore not the primary goal of the integration process. We will therefore use a full text query to the database for each of M entities in the database and produce a short list of length N, N << M of possible matches according to the query results. This will produce a merge list of size <= N*M. Which is a very good result in comparison to the full block merge list M*M.

2.2 Matching

For deciding on the given merge candidate we will try to implement several scoring algorithms and then use the semi supervised approach to create a merging decision function based on scores of the given merge candidate pair. We will use the following two merge candidate scoring algorithms:

- String edit distance algorithm for scoring how close two entity names are in terms of an edit distance [6].
- Cosine distance ranking of the fields associated with the two candidate entities [7].

Once the scores on the merge candidate pair are acquired a semi-supervised learning of the decision function will be implemented similar to [8]. Small set of training data will be acquired either from the users of the IST World portal or by the hand evaluation of the data.

3 OFF-LINE DATA ANALYSIS

The off-line part of the data analysis is performed to compute time and space demanding statistics and competence keywords of the *entities* in either the global context of all the data held in data repository or in the context of a single *entity*. This part of data analysis is therefore performed in advance in the data preparation phase.

The offline data analysis includes (1) supervised learning for automatic classification of the entities (2) graph mining for collaboration graph analysis and (3) multi dimensional scaling for global competence diagram computation. Please note that not all of the here mentioned analysis functionalities have been implemented yet.

3.1 Automatic Classification

We have decided to classify all the entities in the data repository into hierarchical classification scheme called Open Directory Project (DMOZ1).

DMOZ is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of more than 70 000 volunteer editors that actively classify the internet pages into hierarchical categories.

For the automatic classification we used the automatic classifier into DMOZ provided by the TextGarden library [9]. This classifier accepts as input textual description of an entity and outputs the list of most suitable categories for the input text.

For the purposes of the IST World we ran the classifier on all the entities in our database and assigned each entity the one most suitable category according to the constraint list of the allowed and disallowed categories. All together more than 200 000 entities were automatically classified into the science branch of the DMOZ classification hierarchy.

3.2 Entity Importance

We will try to capture the collaborative competence of researchers and research organizations in the context of all the data in the repository which constitutes a big collaboration graph. Scalable graph mining methods will be used to estimate different entity centrality measures [10]. A centrality measure gives an estimate of importance of a node in a social network. There exist many different centrality measures which capture different notions of node importance. We will compute the following centrality measures:

---

[1] The DMOZ page can be found at http://dmoz.org

- *Node Degree*. Node Degree measures the number of neighbours in the social network. It therefore captures the importance of an entity with regards to the process of communication that goes on in a social network.
- *Betweenes*. Betweenes measures the frequency with which a node in a social network falls between pairs of other nodes on the shortest paths connecting them. The measure suggests that an entity is important if it is strategically located on the communication paths linking pairs of others.
- *Closeness*. Closeness measures how close on average a node is to all other nodes in a connected graph. It therefore captures the notion of ease of communication with other nodes.

The centrality measures of the entities will provide different interpretations of the importance of the observed entity. Concern with communication activity will be answered best with a degree-based measure. Interest in control of communication requires a measure based upon betweeness. Finally, a concern on either independence or efficiency leads to the choice of a measures based on closeness.
Several standard algorithms exist for calculation of these measures [10]. We will try to exploit the SQL query language for its computation inside the relational database engine.

3.3 Global Competence Diagram

Competence Diagram [2] is a visualization technique of drawing the entity nodes in a two dimensional plane according to the similarity of their textual description [12]. The Multi Dimensional Scaling of all the entities in the IST World repository will be performed to project the entities from a hyper dimensional space of their textual description onto a two dimensional plane together with the most important keywords and relations between them.
The computed two dimensional coordinates of most important keywords together with the computed two dimensional coordinates of every entity represent the Competence Diagram which can be visualized on a screen. The competence of every entity is described by the position of the entity on the screen and by the relation of the entity to the displayed keywords.

The computation of a global competence diagram is time and space demanding [12]. It will be performed using the TextGarden's singular value decomposition and multi dimensional scaling functionalities as described in part in the Fortuna et al. [12] paper. The functionality is extended by adding the keywords coordinates and relationships. This is achieved by adding artificial documents to the analyzed document collection. The artificial documents are composed according to the singular vectors computed in the process of singular value decomposition. Another problem is scalability of the computation which has to work for more than 100 000 entities. This is resolved using only the acquired artificial documents in the multi dimensional scaling step and then computing the coordinates of the rest of the entities according to their cosine distance to the projected artificial documents. This work on extending existing Document Atlas functionality [12] will be described more thoroughly in the future papers.

4 ON-LINE DATA ANALYSIS

The on-line part of the data analysis is performed to analyze the *entities* in the context of other *entities* currently in the scope of the data selection step as described in section **1**. The analysis in the context of the subset of other entities allows more informative results. E.g. the keywords of an organization's role in a project can be computed according to other participants of the project. Another example is displaying of Collaboration sub Graph of only the authors of the certain paper. As this part of the data analysis depends on the data selected in the selection step, it can therefore not be performed in advance in the data preparation phase and must be done in real time as the user is on line and analyzing the data.
The online data analysis includes (1) Unsupervised learning to calculate the Context dependant Competence Diagram, (2) Unsupervised learning to calculate the clustering of the entities and their time dependant display, (3) Graph drawing algorithm for a nice display of Collaboration Diagrams.

4.1 Context Dependant Competence Diagram

The context dependant competence diagram which we present to our users is a visualization technique similar to the global competence diagram described in section **3.3**. The difference from the global competence diagram is in the selection of the entities to be visualized. The context dependant competence diagram calculates the keywords and the two dimensional coordinates based on the currently selected set entities. This causes the keywords to show the distinction in competence between the visualized entities. This is different form the global competence diagram which would take into an account the textual description of all the entities held in the repository to visualize the selected ones. It would therefore produce descriptive competence analysis. As the competence diagram computation is time demanding only a small subset of less than hundred entities can be analyzed this way in a timely fashion.

The context dependant competence diagram is already implemented. It is built on top of the Document Atlas [12] solution of the TextGarden library [11] as mentioned in the section **3.3**.

4.2 Clustering and Time graph Visualization

The IST World application will use an unsupervised machine learning technique called hierarchical clustering to produce hierarchical groups of selected set of entities which have similar textual description. The intensity of the

groups' activity over time as read from the relational database will then is used to produce an interactive visualization called time graph [13]. This diagram will enable the user an interactive insight into how intensely different research topics were addressed during the past years.

4.3 Graph Drawing

The IST World web data mining application allows the user to visualize the collaboration graph of the selected set of entities. In order for the user to gain an insight into the collaboration patterns of the selected entities the graph has to be nicely visualized. We use the TextGarden's graph visualization utility [14] to perform graph visualization optimization in real time. The algorithm presents the drawing problem as an optimization problem, which is then solved using the gradient descent algorithm. The collaboration diagram is computed using a single SQL query to the relational database engine.

5 CONCLUSIONS

We are developing a data mining application that will allow its users to gain insight into competence and collaboration details of the individual researcher, research organizations, research projects and publications. In this paper we gave an overview of the different machine learning and data mining approaches for solving some of the problems of implementation of the system. As the data in the IST World repository has to be integrated, full text querying, string matching and semi supervised learning methods will be used to tackle the problem of blocking and matching. All the context independent data analysis is performed in advance. This includes automatic classification into DMOZ hierarchical classification scheme, graph mining to estimate different measures of importance of entities and global competence diagram for evaluating the competence of the entities in the scope of all the entities. The context dependant data analysis must be performed online. This includes context dependant competence diagram, clustering of the selected set of entities and graph drawing optimization algorithm. Not all of the approaches described in this paper have been implemented yet.

Acknowledgments

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under IST-World (FP6-2004-IST-3-015823).

References

[1] Jörg B., Ferlež J., Grabczewski E, Jermol M. IST World: European RTD Information and Service Portal. 8th International Conference on Current Research Information Systems: Enabling Interaction and Quality: Beyond the Hanseatic League (CRIS 2006), Bergen, Norway, 11-13 May 2006

[2] Ferlež J. Public IST World Deliverable 2.2 – Portal Services and Functionalities Definition. http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D2.2_PortalServicesAndFunctionalitiesDefinition.pdf

[3] Jörg, B., Jermol M., Uszkoreit H., Grobelnik M., Ferlež J.; Analytic Information Services for the European Research Area. To appear In Proceedings: eChallenges2006 e-2006 Conference, October 25-27, 2006. Barcelona, Spain.

[4] Grabczewski E. Public IST World Deliverable 3.2 – Base Set of Data http://ist-world.dfki.de/downloads/deliverables/ISTWorld_D3.2_BaseSetOfData.pdf

[5] Winkler E., Overview of Record Linkage and Current Research Directions. RESEARCH REPORT SERIES (Statistics #2006-2)

[6] Koudas, N., Marathe, A., and Srivastava, D. (2004), "Flexible String Matching Against Large Databases in Practice," Proceedings of the 30th VLDB Conference, 1078-1086.

[7] Newcombe, H.B. and Kennedy, J. M. (1962) "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information" Communications of the Association for Computing Machinery, 5, 563-567.

[8] Winkler, W. E. (2001), "The Quality of Very Large Databases," Proceedings of Quality in Official Statistics '2001.

[9] Grobelnik M., Mladenic D. Simple classification into large topic ontology of Web documents. In Proceedings: 27th International Conference on Information Technology Interfaces (ITI 2005), 20-24 June, Cavtat, Croatia.

[10] Freeman L., Centrality in Social Networks Conceptual Clarification, Social Networks, 1 (1978/79) 215-239

[11] TextGarden software suite, http://kt.ijs.si/Dunja/textgarden/

[12] Fortuna B, Mladenič D, Grobelnik M. Visualization of text document corpus. Informatica journal, 29(4):497–502, 2005.

[13] Shaparenko B., Caruana R., Gehrke J., Joachims T., Identifying Temporal Patterns and Key Players in Document Collections. Proceedings of the IEEE ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications (TDM-05), 165–174, 2005.

[14] Brank, J. Drawing graphs using simulated annealing and gradient descent. Zbornik C 7. mednarodne multi-konference Informacijska družba IS-2004, 67-70.