ECT 584: Project

Students can choose to do either an implementation project, a research paper, or a data analysis project. A project proposal (1 page) must include the project type (one of the 3 categories below) and…

- Implementation: project description, proposed methodology, techniques, approaches, implementation choices, resources used, and the tentative schedule.
- Research Paper: abstract, a short (but detailed) outline, and a list of reference sources to be used in the research.
- Data analysis project: detailed description (and samples) of the data to be analyzed, the data mining problems to be solved, the techniques to be employed to solve the problems, and the tools to be used.

Implementation Projects

The implementation projects may involve implementing and performing experimental evaluation on one or more techniques discussed in the course (e.g., clustering, association rules, classification, etc.), or combining various data mining techniques (possibly using available tools) into a Web data mining solution for a specific problem.

Some specific examples of implementation projects include:

- Implementing or extending one of the data mining techniques discussed in class (or related techniques) and testing the implementation on various test data sets, such as Web usage or e-commerce data or other test data sets available from the UCI KDD Archive.
- Implementing or extending the techniques discussed in class for preprocessing of Web usage data, including user/session identification, path completion, automatic discovery and filtering of robot navigation, and pageview identification. Ideally, this should be implemented as an extension of the WEKA data mining package, so that the results of preprocessing can be directly used as input for DM components.
- Designing and implementing a data warehouse for integration and management of Web usage, structure, content, and e-commerce data, and analyzing this data by performing OLAP queries against the data warehouse, and using the results as input to data mining algorithms.
- Implementing a system to analyze the effectiveness of a Web site by comparing the site structure to the navigational behavior of users, analyzing site and user e-metrics, and predict user behavior for individual or segments of users.
- Implementing a recommender system based on usage mining, content filtering, or collaborative filtering techniques discussed in class.
- Designing an automated classification tool that uses machine learning to automatically identify and classify robot navigation sessions from Web usage log files.
- Designing a query language for querying interesting rules or patterns resulting from Web usage or Web content data.

- Developing your own mail, news, or Web information filtering agent (e.g., an agent that extracts information about a particular topic/product from specific sites). The agent design must include one or more machine learning and data mining techniques such as classification, clustering, association rule mining, Markov models, etc.

Research Papers

Research papers involve doing an in-depth study, survey, or evaluation of one or more topics related to Web data mining. A research paper may examine the use of specific data mining or Web mining techniques in one or more application areas. A research paper must relate to one or more of the topics discussed in class, but must not be simply a summary of the material covered in class or the readings. The goal of such a "research project" is to go beyond the class material and examine one of the topics in a much more in-depth manner. The evaluation of the papers will be based on thoroughness (including adequate coverage of relevant issues/techniques as well as references to related work), soundness (including justification for any claims made, illustrative example, correct and adequate analysis of connections or relationships among concepts or techniques), clarity/organization, and significance (defined by the degree to which the paper covers new material, and the extent of the original work by the author in drawing conclusions and synthesizing scholarly work in this area). Note that a research paper must not simply be a concatenation of material from several other papers, but must include some original analysis of that work in the context of the paper topic. The suggested length for a research paper is 15-20 single-spaced pages.

Some examples of topics for research paper topics are:

- The applications of Web usage mining in various areas (e.g., for e-commerce, user profiling, Web personalization, etc.). Note: The paper should contain material and sources substantially different than, and go beyond, the material covered directly in class or in readings.
- Recommender Systems and User Profiling: a comparative study of various techniques in data mining and collaborative filtering to learn user profiles and predict future user behavior. The study should examine variety of techniques and approaches used in the design of recommender systems (including collaborative filtering, content-based filtering, model-based approaches which use data mining, and hybrid systems).
- Web content mining/Text mining: a study of various techniques to mine information and patterns from semi-structured such as text, or Web documents. The paper should examine topics such as the applications of text mining on the Web, information extraction and mining of data records from the Web, concept discovery from document collections, document categorization and classification, etc.
- Web structure mining: a study of various techniques to mine knowledge from the linkage structure of the Web. Among the topics that can be explored in this study are application of structure mining in information retrieval (such as Google's Pagerank algorithm), algorithms based on the notions of Hubs and Authorities, and the automatic discovery of Web communities.
- Data Mining on the Social Web: Social Web (Web 2.0) technologies allow users to connect, share resources, and actively generate content on Web sites. How can the rich data in such social

networking sites such as Facebook and Twitter, or in resource sharing sites such as Flickr, Last.fm, and Delicious be mined to help users interact with these sites more effectively. Among the topics that can be explored are the use of Social Network Analysis algorithms to discover interesting relationships, using data mining and machine learning algorithms to predict user behavior or to recommend resources and users (friends), and application for tag suggestion/ recommendation in social tagging Web sites.

- Web Data Warehousing: detailed study of Web data warehouses and datamarts, and their use, along with various data mining techniques for decision support and gaining business intelligence from Web data. Note: The main focus of this study must not be Data warehousing, but rather their use in the context of Web and e-commerce data, as well as their connections to data mining or data analysis.
- Web Information Agents: a study of the use of client-side (or server-side) agents on the Web that assist users in filtering information and in browsing or searching tasks on the Web.

Data Analysis Projects

Data Analysis projects involve the application of data mining and Web mining techniques discussed in class or in readings to one or more specific data sets. The goal of DA projects is to go through the full data mining cycle with respect to a particular data set (including the specification of the business problem to be solved, the specification of the data mining tasks to be performed, selection, preprocessing, integration, and transformation of the data, application of several DM tasks and the discovery of patterns, evaluation of patterns, and recommending specific actions with respect to relevant findings). A DA projects may involve the application of data mining in a particular domain and data set with which you are familiar such as your work, the Web, e-commerce, etc. The final report should include a detailed analysis of the complete scenario for the application of the KDD process, including specification of the DM problem (based on application objectives), data collection, data preparation, pattern discovery using a variety of data mining and statistical techniques, interpretation of results, and conclusions).

Some examples of data analysis projects include:

- Performing data mining on Web usage (or e-commerce) data from a particular Web site in order to analyze the behavior of users, including various site metrics, user metrics, user segments, associations, and opportunities for personalization. The project plan must include all aspects of Web usage preprocessing.
- Performing the full KDD cycle of a real data set (other than Web usage or e-commerce data). Examples of such data may include (but are not limited to) user or customer data from a real business or organization, census or demographic data, sensor data for device diagnostics, network traffic data, music playlist data obtained from music sharing Web sites, social networking data (such as social tags obtained from sites).
- Examination of one or more specific commercial or freely available data mining packages, other than those used in class. In this option, you must be able to install and experiment with the system. The package must be compared in detail with other comparable products. The final report must include a technical evaluation and provide a critical analysis of the results of applying various

KDD capabilities provided by the software on at least two realistic data sets, such as test data sets available from the UCI KDD Archive.

Final Project Checklist

The following is a final checklist of what you need to turn in for your project. Please note that these are general requirements and in specific cases, not all of these requirements may apply. If you are not sure about what to include, please consult with me before your submission. Also, if you think that installation of your system may require non-trivial or unusual steps, you can arrange for a demonstration of your system.

Implementation Projects

The Implementation projects will be evaluated based on significance, design, correctness, documentation, and appropriate evaluation/testing. You will need to electronically submit a compressed file containing your project distribution files and documentation. Your project documentation should contain the following components:

A detailed description of your system (including specific techniques and algorithms you used), and the interaction between the components (make references to code segments, modules, methods, functions, etc. as necessary). Your write up should also include an evaluation of your system demonstrating its correctness and functionality. If you used any outside sources in your implementation, please clearly indicate which sources, and how and where they were used.

Complete (actual) sample runs of your program with description, illustrating how your system works, along with any intermediate input or output used for the sample run.

Your project distribution files should contain the following:

- Complete source code (be sure that your source code is fully documented and easy to read).
- Binary files (e.g., executables, DLLs, Class files) or other components necessary to run your program.
- Readme file containing instructions on how to compile, install, and/or run your program.
- Project report/documentation, as described above.
- Any test data used for evaluation of your system.
- If your application is CGI-based or otherwise has a server component, please provide a URL for a demo version of your system.

Research Papers

Research Papers will be evaluated based on thoroughness, soundness, clarity and organization. The overall structure of the paper is up to you, but you must have the following sections in addition to the main body of the paper:

- Abstract: This is a short synopsis of the main points of the paper. This should be 200-300 words, and should appear along with the title and your name, ID number, and email, on the first page. The rest of the paper should start on page 2.

- Conclusion: Summarize your conclusions and findings. You conclusion section must include your own original thoughts and discussion of your findings.
- References: This is a list of references that you have used in doing your research and throughout your paper. The references should be numbered and the number for the reference should appear in the appropriate places in the text of the paper where the reference was used (it is not enough to list a bibliography at the end of the paper without actually using any references within the body of the paper). You can look at any of the papers assigned for reading in the class for acceptable uses of references. URL references should only be used for referring to specific system Web sites and not as a way to reference published work.

It is highly recommended that you create appropriate sections and subsections to create an easily understandable organization corresponding to the topics covered. Your final paper should be submitted electronically in PDF format. Submissions in other formats will not be accepted.

Data Analysis Projects

Data analysis projects will be evaluated based on significance, design (specification of the data mining problems to be used), thoroughness (completeness of the analyses and DM tasks performed on the data, as well as well as thorough discussion of the findings and conclusions), and documentation. You will need to electronically submit a compressed file containing your project distribution files and your project report. Your project report should contain the following components:

- An executive summary summarizing the project goals, methods used, and conclusions.
- The main body of the report providing details on the data set used, design decisions, description of how the KDD cycle was applied, tools used to perform specific DM tasks, a detailed description of the results of data mining, and conclusions.

Your project distribution files should contain the project report (as described above), any code written to perform specific tasks or to prepare the data, the full data set (or a sufficiently descriptive sample of the data) used for the project.