A Sample "SciBot" Project Report

This is a skeleton of a report on a complete but least-effort project. For every task, there are ways to go beyond the minimum effort. The key is to explain what you did that was beyond the minimum and also explain whether that additional work generated better results.

1. Introduction/Motivation

A reasonable problem statement, data domain description, data set(s) description, overview of tasks and workflows used to perform tasks, and preview of conclusions should go here.

2. Approach

2.1. T1: Data Cleaning

We consolidated all the data files into one file/database. We carefully checked if the dataset is incomplete, noisy, or redundant. All issues were resolved. Here are statistics of the dataset:

Number of papers: NNN
Number of authors: MMM
Number of venues: XXX
… other statistics …

2.2. T2: Entity name detection (feature extraction from paper)

We used functions in paperclassification.py (available at http://www.meng-jiang.com/teaching/paperclassification.zip).

The Abbreviation rule (Hint 3 on the project handout) was used to find quality entity names. We use absolute count (of the name appearing together with its abbreviation) to measure the quality of the names. Here are the top 10 entity names and their quality scores (see phrase2count.txt).

Entity name	Quality
Latent dirichlet allocation	240
Support vector machine	213
Support vector machines	208
… seven more rows …	

We randomly select 30 from the 200 entity names with highest quality and label them as "entity

name" and "non-entity name". The accuracy is xxx.

[use other hints in the handout to go above and beyond the minimum!]

2.3. T3: Entity typing (feature-to-dimension classification)

We used functions in entitytyping.py (http://www.meng-jiang.com/teaching/TypingDemo.zip).

For each entity name discovered in T2, we determine whether it identifies a Problem, Method, Metric, or Dataset by counting the number of occurrences of "trigger" words for each possible type, with the highest count identifying the type. Here are the top 10 entity names of the highest quality score, and their types (see entitytyping.xlsx, find corresponding lines). Note that the count is the absolute count.

ENTITY NAME	COUNT	TYPE
latent dirichlet allocation	775	Method
support vector machine	719	Method
support vector machines	1385	Method
...		...

We randomly select 30 from the top 200 entity names and label them as one of the four types. The accuracy is xxx.

[consider using other methods like ensembles to generate different/better typing results!]

2.4. T4: Collaboration discovery (frequent pattern mining)

Here each paper is considered as an itemset and each author is an item. We use Apriori to find frequent itemsets (i.e., frequent author collaborations). *We use functions in ResponseBotDemo.py (http://www.meng-jiang.com/teaching/ResponseBotDemo.zip).* We set the min_sup as 3. We find xxx frequent itemsets in total. Here are the 10 most frequent itemsets (collaborations).

AUTHORSET (ITEMSET)	SUPPORT
...	...

[try using FP-Growth or find closed itemsets to go beyond the minimum!]

2.5. T5: Problem-to-method associations (association mining)

Here we use association mining methods to find associations X→Y where X is a Problem item, Y

is a Method item. We evaluate the support, confidence and lift measure of each association. *Again, we use functions in ResponseBotDemo.py:*

(http://www.meng- jiang.com/teaching/ResponseBotDemo.zip).

Here are the top 10 associations of the highest ***support***.

ASSOCIATION	**SUPPORT**	CONFIDENCE	LIFT
"recommender systems"→"matrix factorization"			
...			...

Here are the top 10 associations of the highest ***confidence***.

ASSOCIATION	SUPPORT	**CONFIDENCE**	LIFT
"recommender systems"→"latent dirichlet allocation"			
...			...

Here are the top 10 associations of the highest ***lift measure***.

ASSOCIATION	SUPPORT	CONFIDENCE	**LIFT**
"recommender systems"→"root canal without anesthesia"			
...			...

Our observation is...

[consider mining other kinds of associations like Dataset-to-Problem (what problems have been often studied using a specific data set?)]

2.6. T6: Paper-to-conference classification

Here we only use the top xxx (maybe 100) quality entity names as features of papers. Each paper is a data object. Each entity name is a binary feature (if the paper has this entity name or not). We set up a binary classification task: we classify papers into positive (a KDD paper) and negative (a non-KDD paper). As we only use the top xxx entity names and we only consider papers that have at least xxx (maybe 3), we have only xxx (maybe 500) papers: we have xxx (maybe 120) positive and xxx (maybe 380) negatives.

We use ID3 to build a Decision Tree model for paper to classification. The height is limited to 10,

i.e., the top 10 informative features/entity names. We did hold out validation for 10 times: 80% for training and 20% for testing. We evaluate the performance using Accuracy, Precision, Recall, and F1. Here is the result: ...

[Go above and beyond the minimum! Consider using either numerical feature (e.g., the number of entity names in a paper) or PCA to reduce the number of features; consider using other classification methods (like C4.5, CART, Naïve Bayes, Neural Networks, SVM) to solve the problem; consider applying other evaluation methods (like 10-fold cross-validation) or any other evaluation metrics (like ROC curves, MAP, MAE, RMSE); consider doing multi-class classification to take advantage of the fact that the data set has four conferences.]

2.7. T7: Paper clustering

We use the same data objects and features used in T6 for paper clustering. In other words, we group 500 papers based on 100 binary features into four groups. We first use PCA to generate 10 numerical features and then use K-Means clustering to solve the problem. Here is the result... (visualization) We evaluate the performance using...

[To infinity and beyond! Develop a proximity measure for papers based on the entities, and cluster papers using a hierarchical technique! Or use the features and try out Kernel K-Means, K-Medoids, K-Medians or some other convenient method.]

2.8. Additional tasks

[Up to you! Visualize something, or write a random paper generator using entity names, or …]

III. Conclusion

A summary of the work and the conclusions that were reached. Comment on how this project increased your knowledge of data mining and/or your level of skill with data mining techniques.