

Data Mining Project – Spring 1394

This is an exploratory project. You are encouraged to collect interesting data sets for an application domain that interests you. The more data you collect the better it is for finding interesting patterns. Your project should consist of the following three phases:

Data Collection (for the domain you like/are interested in)

Data Preprocessing (data cleaning and transformation into a useful form)

Data Mining (using algorithms you have seen so far)

At least one of these phases should be not trivial. For example the data collection and preprocessing phase could be non-trivial (e.g. the sites you use have some specific APIs that you need to use and/or the data needs special preprocessing).

Or the data mining process could be non-trivial. E.g. you collect a lot of data and then some algorithms such as those in Weka, being main memory algorithms, have a hard time to mine your data. In such a case, you should modify or recode those algorithms, or research available algorithms that can be more efficient for large amounts of data.

You should present your work in person and submit a report (in English) describing the main points of your project and results.

You should write the report in a research paper format. See <http://infolab.stanford.edu/~widom/paper-writing.html> on how a paper can be structured.

Some interesting data/articles references are:

<http://www.kaggle.com/c/titanic-gettingStarted>

<http://archive.ics.uci.edu/ml/datasets/YouTube+Comedy+Slam+Preference+Data#>

<http://archive.ics.uci.edu/ml/datasets/Farm+Ads>

<http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD>

<http://archive.ics.uci.edu/ml/datasets/URL+Reputation>

<http://cseweb.ucsd.edu/~voelker/pubs/mal-url-icml09.pdf>

<http://infolab.stanford.edu/~echang/recsys08-69.pdf>

http://www.yelp.ca/academic_dataset

<http://www.grouplens.org/node/73>

<http://2013.msrconf.org/challenge.php>

Also, the practical book:

Programming Collective Intelligence by Toby Segaran
contains a wealth of ideas on how to find data and how to mine them.
You can search for the pdf online.