Text mining algorithms for large data sets

Josef Bušta

18. 1. 2013

Text mining tasks, text analysis

- text categorization, text clustering, concept/entity extraction, document summarization, entity relation modeling
- information retrieval, pattern recognition, tagging/annotation

Online machine learning

Given a loss function L and a stream of examples S of the form (x,y), do the following

- Initialise a starting model w
- While there are more examples in S
  - Get the next feature vector x
  - Predict the label y' for x using the model w
  - Get the true label y for x and incur a loss L(y,y')
  - Update the model w if $y \neq y'$

Online machine learning: algorithms

- Large-scale and sparse date
- Implementations:
  - Opal (http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/opal/),
  - ARROW++ (http://code.google.com/p/arowpp/),
  - Oll (http://code.google.com/p/oll/),
  - SVMlin (http://vikas.sindhwani.org/svmlin.html),
  - LIBLINEAR (http://www.csie.ntu.edu.tw/~cjlin/liblinear/)

Simple Iterative Multiple Projection on Lines (SIMPLE)

- http://www.cse.iitb.ac.in/soumen/doc/vldbj2003/simpl2003.pdf
- Algorithm for text classification via multiple linear discriminant projections
- Nearly linear-time classification (in the number of terms (dimensions) plus the number of documents)

Simple Iterative Multiple Projection on Lines (SIMPLE)

(1) Find a series of projections of the training data by using Fisher's linear discriminant as a subroutine

(2) Project all training instances to the low-dimensional sub-space found in the previous step

(3) Induce a decision tree on the projected low-dimensional data

Fisher's linear discriminant is a (unit) vector α such that the positive and negative training instances, projected on the direction α, are as "well-separated" as possible.