# Text Classification using Wordnet

Pravin Paratey

May 08, 2007

# What is classification?

- Classification (Tagging) is assigning one or more categories (*tags*) to documents, based on its contents.
- DMOZ (ODP), Yahoo Directory, Technorati, Del.icio.us, StumbleUpon

# Need for classification

- Automatically construct hierarchies from unstructured data. eg. WWW, Web Forums, Local data on disk
- Spam filtering (Email/Web)
- Ranking aggreggated data (RSS/Atom) according to user preferences. Showing only what user is likely to read.
- Search Engines to better rank the document
- Create Tag Clouds to analyze data

# Exisiting Classifiers

- Naive Bayes
- TF-IDF
- Latent Semantic Indexing
- Support Vector Machines
- Artificial Neural Network
- k-NN
- Decision Trees
- Concept Mining

# Approach

- POS-Tagging the document

# Approach

- POS-Tagging the document
- Stopword removal

# Approach

- POS-Tagging the document
- Stopword removal
- Constructing Synset Map
  - For each word, create an entry only if it is a noun or a verb

# Approach

- POS-Tagging the document
- Stopword removal
- Constructing Synset Map
  - For each word, create an entry only if it is a noun or a verb
- Analyze Hypernymy
  - For each word, traverse its Hypernymy relation upto a depth x (I took x = 2)
  - Assign maximum score to the current Synset and reduce score with distance

# Approach

- POS-Tagging the document
- Stopword removal
- Constructing Synset Map
  - For each word, create an entry only if it is a noun or a verb
- Analyze Hypernymy
  - For each word, traverse its Hypernymy relation upto a depth x (I took x = 2)
  - Assign maximum score to the current Synset and reduce score with distance
- Output likely tags
  - Likely tags are the ones with maximum score
  - Output their noun forms

# Demo

- Demo of DocTagger

# Conclusion

- Wordnet is a powerful tool to use for classification
- Exploiting Hypernymy relation increases correctness by a large factor

# Future Work

- Use Meronymy/Holonymy relations
- Proper Noun resolution
- For words with multiple senses (eg. Mouse: *rodent*, *computer*), use similarity measure using the sentence
- Run on an extended corpus to gain correctness statistics (eg. Wikipedia)

# References


- C. Fellbaum. Wordnet: An Electronic Lexical Database. MIT Press, 1999
- Li, Y.H. and Jain, A.K. Classification of text documents. Computer Journal, 41 (8) , p. 537-46, 1998.
- Scott, Sam and Stan Matwin. Text classification using WordNet hypernyms. Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998.
- A. Hotho, S. Staab, and G. Stumme. Wordnet improves text document clustering. In submitted, 2003. http://citeseer.csail.mit.edu/hotho03wordnet.html
- Vlajic, N. and Card, H. C. Categorizing Web pages using modified ART. Canadian Conference on Electrical and Computer Engineering, Vol.1, p. 313-316, 1998.

# References (cont ...)


- Jing, Hongyan and Tzoukermann, Evelyne. Information retrieval based on context distance and morphology. SIGIR '99, Proceedings on the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 90-96, 1999.
- Zamir, O. and Etzioni, O. Web document clustering: a feasibility demonstration. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 46-54, 1998.
- Douglass Cutting, David R. Karger, Jan O. Pedersen and John W. Tukey. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections. In Proceedings of ACM/SIGIR, p. 318-329, 1992.