

## **Selected references for Text Mining**

### **Named Entity Extraction**

Nadeau & Sekine, A survey of named entity recognition and classification *Linguisticæ Investigationes* 30:1 (2007)  
- a recent survey

Andrew McCallum, Wei Li. 2003. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. In: *Proceedings of CoNLL-2003*, Edmonton, Canada. 188-191.

- early paper in a long series on using CRFs for named entity recognition

Indrajit Bhattacharya and Lise Getoor, Collective Entity Resolution in Relational Data *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1(1), 2007

### **Machine Learning methods**

John Lafferty, Andrew McCallum, Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proc. of 18th Int. Conf. on Machine Learning*.

- CRFs are the most widely used method in NER

Cohen, W. AND Richman, J. 2002. Learning to match and cluster large high-dimensional data sets for data integration. In *The ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. Edmonton, Canada.

- application to schema matching and data integration

### **Information Extraction**

Andrew McCallum, *Information Extraction*, ACM Queue 2005.

- good introduction with references

X. Li, P. Morie, and D. Roth, Semantic Integration in Text: From Ambiguous Names to Identifiable Entities. *AI Magazine. Special Issue on Semantic Integration (2005)* pp. 45-48

- entity resolution in databases

Ray, S. and Craven, M. Representing sentence structure in hidden Markov models for information extraction. in *Proceedings of the 17th International Joint Conference on Artificial Intelligence*. 2001.

Grishman, R., The role of syntax in Information Extraction, in *Advances in Text Processing: Tipster Program Phase II*. 1996, Morgan Kaufmann.

- old but classic

### **Open IE and Relation extraction**

Etzioni, O., et al., Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence*, 2005. 165(1): p. 91-134.

Agichtein, E., and Gravano, L. 2000.

Snowball: Extracting Relations from Large Plain-Text Collections. in *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 85-94. San Antonio, TX: Association for Computing Machinery.

Oren Etzioni, Michele Banko, Michael J. Cafarella, *Machine Reading*, extended version; AAAI, 2006

Michele Banko and Oren Etzioni

The Tradeoffs Between Open and Traditional Relation Extraction

*Proc. 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*

Textrunner <http://www.cs.washington.edu/research/textrunner/>

SRES,

### **Sentiment analysis and Market Structure analysis**

Pang, B., Lee, L., and Vaithyanathan, S., Thumbs up? Sentiment Classification using Machine Learning Techniques, in Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing. 2002, Association for Computational Linguistics, Morristown, US: Philadelphia, US. p. 79-86.  
- seminal paper

Dave, K., Lawrence, S., and Pennock, D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. in Proceedings of the Twelfth International World Wide Web Conference WWW-2003. 2003.

Hu, M. and Liu, B. Mining and summarizing customer reviews. in KDD-2004. 2004.

Popescu, A.-M. and Etzioni, O. Extracting Product Features and Opinions from Reviews. in Proceedings of HLT-EMNLP. 2005.

Esuli, A. and Sebastiani, F. Determining the semantic orientation of terms through gloss classification. in Proceedings of CIKM-05. 2005. Bremen, DE.

Kim, S.-M. and Hovy, E. Automatic Identification of Pro and Con Reasons in Online Reviews. in Proceedings of the Conference on Computational Linguistics/Association for Computational Linguistics (COLING/ACL-2006). 2006. Sydney, Australia.

Feldman, R., Fresko, M., Goldenberg, J., Netzer, O., Ungar, L.:  
Extracting Product Comparisons from Discussion Boards.  
Proceedings of ICDM-2007, Omaha, NE. (2007)

Lise Getoor Link mining: a new data mining challenge  
ACM SIGKDD Explorations Newsletter 84 - 89, 2003

### **Visualization**

Batagelj, V. and Mrvar, A. Pajek - Analysis and Visualization of Large Networks. in Graph Drawing Software. 2003. Berlin: Springer.

Gregory, M., et al. User-Directed Sentiment Analysis: Visualizing the Affective Content of Documents. in Sentiment and Subjectivity in Text Workshop at the Annual Meeting of the Association of Computational Linguistics (ACL 2006). 2006.

Feldman, R., Kloesgen, W., and Zilberstein, A. Visualization Techniques to Explore Data Mining Results for Document Collections. in Proceedings of KDD '97. 1997.

### **Scalable algorithms**

Jeffrey Dean and Sanjay Ghemawat,  
MapReduce: Simplified Data Processing on Large Clusters,  
OSDI'04: Sixth Symposium on Operating System Design and Implementation, 137-150, 2004,

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz and Yoram Singer  
"Online Passive-Aggressive Algorithms", Journal of Machine Learning Research", 7, 551-585, 2006.

Abhinandan Das, Mayur Datar, Ashutosh Garg, Shyam Rajaram  
Google News Personalization: Scalable Online Collaborative Filtering

WWW, 2007

S. Guha, N. Mishra, R. Motwani, L. O'Callaghan, Clustering Data Streams,  
IEEE Symposium on Foundations of Computer Science., 2000