Master of Technology in Knowledge Engineering

Text Mining

Clustering

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
email: zhenzhen@nus.edu.sg

Agenda

- Clustering in general
  - Similarity measures
  - Clustering algorithm
    - Hierarchical clustering
    - Partitioning clustering
- Clustering in text mining
  - Document clustering
  - Word clustering
  - Cluster Visualization

Objectives

- To give a brief overview of clustering techniques
- To introduce two major types of clustering in text mining – document clustering and word clustering

Clustering in General

Clustering

- *Clustering*, or cluster analysis, is the process of automatically identifying similar items to group them together into clusters.
  - *Unsupervised learning* –no labeled training examples need to be supplied; no prior knowledge of the number of groups,
  - Originated in the fields of statistics and data mining, used on numerical data
- A good clustering method will produce high quality clusters in which
  - Items in the same cluster are very similar to each other.
  - Each item is less similar to items in other clusters.

Clustering Algorithms

- Components needed for a clustering algorithm:
  - A method for computing similarity between items
  - An efficient method for comparing all of the items to be clustered.
- Some algorithms may require the number of clusters to be found as input
- The similarity measure (based on distance functions) depends on the type of data.

Similarity measures

- Similarity measures for **structured** data
  - Binary data
    - E.g. Hamming Distance – the number of attributes for which the corresponding values from the two items are different

Item A: 1 0 1 0 1 1 1
Item B: 0 0 1 1 0 1 0

$d_{AB} = 4$

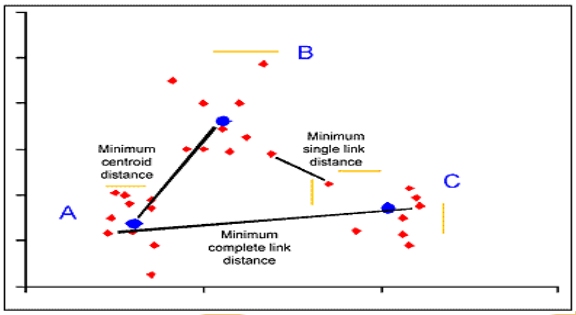
  - Numerical data
    - E.g. Euclidean Distance

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

    - Typically normalization is applied first

Distance between Two Clusters

- ***Single link method***- The distance between two clusters is equal to the distance between the two closest records in them.
- ***Complete link method***- The distance between two clusters is equal to the distance between the two most distant records in them.
- ***Centroid method-*** The distance between two clusters is equal to the distance between their centroids.

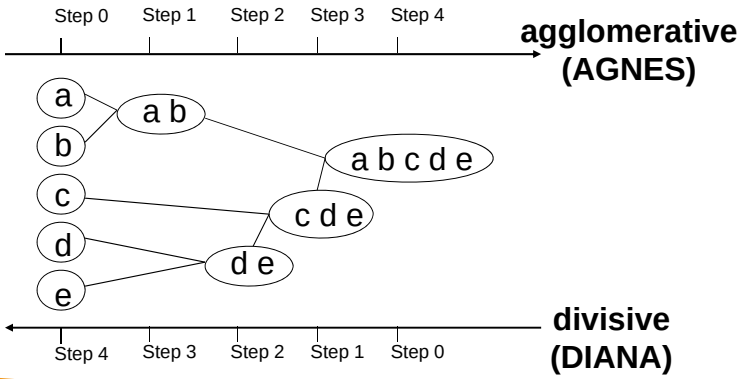


Major Clustering Algorithms

- **Hierarchical Clustering**
  - Iteratively groups documents into cascading sets of clusters.
  - Top-down (divisive) approach – Items are split iteratively based on their similarity measures.
  - Bottom-up (agglomerative) – Items are joined together iteratively.
- **Partitioning Clustering**
  - Constructs various partitions and then evaluate them by some criterion
  - Most popular type – k-means and its variants (k-medoids and k-medians)
- **Spectral clustering**
  - using spectrum of the similarity matrix to perform dimensionality reduction before clustering in fewer dimensions

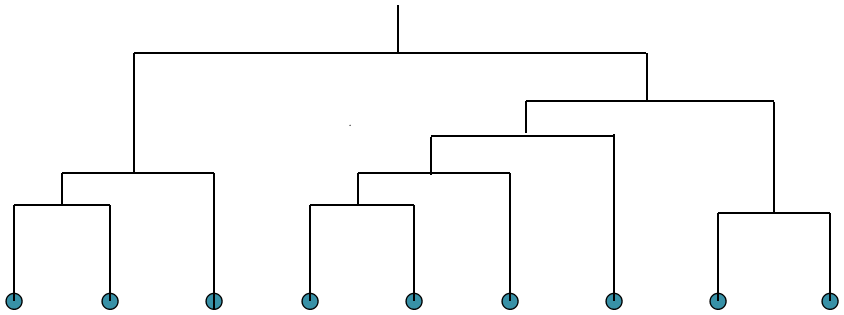
Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition



Dendrogram

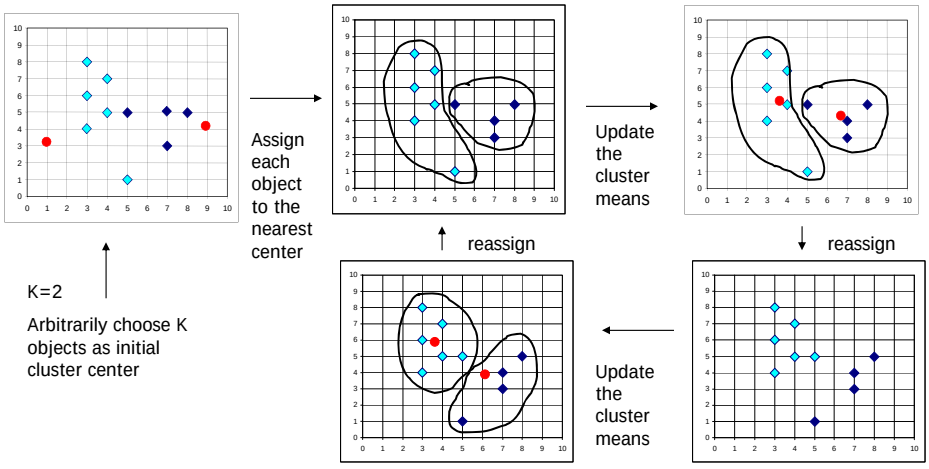
- Clustering process leads to several levels of nested partitioning (tree of clusters), called a *dendrogram*.
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



Partitioning Clustering

- Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means*, *k-medoids*, and *k-median* algorithms
  - ***k-means*** : Each cluster is represented by the center of the cluster calculated using mean.
  - ***k-medoids*** or PAM (Partition around medoids): Each cluster is represented by one of the objects(most centrally located) in the cluster
  - ***k-median***: The center of the cluster is calculated using median

The *K-Means* Clustering Method



Iteration until there's no more new assignment

Clustering in Text Mining

-Clustering similar documents
- Clustering similar words

Document Clustering

Document Clustering vs. Text Classification

- Document/Text Classification (*supervised* learning)
  - Looks at stored examples with correct answers and projects answers for new examples.
  - The answers, or predetermined class labels, must be available.
- Document Clustering (*unsupervised* learning)
  - Groups together documents with similar content into the same cluster.
  - The number of the clusters and their labels are not known before clustering.
  - Ideally each document is very similar to the other documents in its cluster and much less similar to documents in other clusters

Applications of Document Clustering

- Especially beneficial for exploratory analysis of textual data
- Document organization and browsing
  - the hierarchical organization of documents into coherent categories for systematic browsing of the document collection
- Corpus summarization
  - to provide summary insights into the overall content of the collection in the form of cluster-digests or word-clusters
  - Applied in various domains as an inexpensive way to summarize and organize documents, e.g., grouping problems reported to help desks
- As a pre-processing step for document classification

Document Clustering

- It also requires converting the *unstructured* text in each document into a *structured* representation before applying the clustering algorithm.
- The most popular representation is vector space, where each document is represented as a vector noting the words appearing in the document.
- Similarity between two documents are measured using a vector distance metric, e.g. *cosine similarity*, *Jaccard's coefficient*, etc.
- It can be combined with a hierarchical or partitioning method.

Cosine Similarity

- A similarity measure between two vectors by measuring the cosine of the angle between them

$$Sim(D_i, D_j) = \frac{D_i \bullet D_j}{|D_i| * |D_j|} = \frac{\sum_k w_{ki} w_{kj}}{\sqrt{\sum_k w_{ki}^2 \sum_k w_{kj}^2}}$$

- Example: Given 3 document vectors shown here

$|D_1| = \sqrt{0.1761^2 + 0.4771^2 + 0.1761^2} = \sqrt{0.2896} = 0.5382$

$|D_2| = \sqrt{0.4771^2 + 0.4771^2 + 0.1761^2 + 0.1761^2} = \sqrt{0.5173} = 0.7192$

$|D_3| = \sqrt{0.1761^2 + 0.4771^2 + 0.9542^2 + 0.1761^2} = \sqrt{1.2001} = 1.0955$

$Sim(D_1, D_2)$ = (0.1761*0.1761)/(0.5382*0.7192)=0.0801

$Sim(D_1, D_3)$ = (0.4771*0.9542+0.1761*0.1761)/(0.5382*1.0955)=0.8246

D_1	D_2	D_3
0	0	0
0	0	0.1761
0	0.4771	0
0	0	0.4771
0	0.4771	0
0.1761	0.1761	0
0	0	0
0	0	0
0.4771	0	0.9542
0	0.1761	0
0.1761	0	0.1761

Dimensional Reduction

- With big document collections, the dimension of the vector space may easily range into tens of thousands.
- Feature selection is very important for performance reasons.
  - Many good feature selection methods available for classification, using the distribution of features in classes as found in the training documents.
  - Such distribution is not available in clustering.
- Alternative approach – dimension reduction
  - By mapping a high-dimensional feature space to a much lower dimensional subspace
  - E.g. *Latent Semantic Indexing* or *Singular Value Decomposition*, etc.

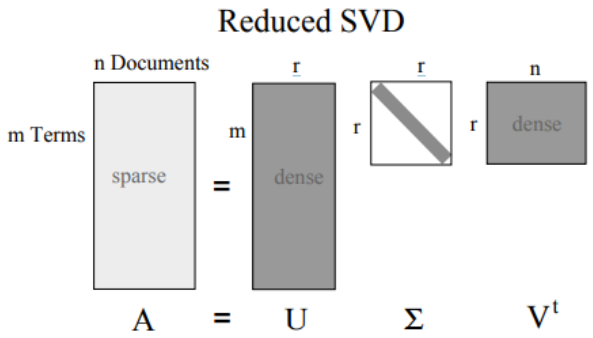
Singular Value Decomposition

- SVD reduces the dimensionality of a data matrix by calculating linear combinations of existing variables
  - Each successively constructed linear combination of variables extracts from the data matrix the maximum amount of "information".
  - The linear combinations of variables are orthogonal to (or independent of) each other, so each linear combination contains " different information".
- Particularly useful in text mining and statistical natural language processing
  - Operating on the term-document matrix
  - Approximating the original matrix, maximizing the information extracted from that matrix

Singular Value Decomposition

- Objective similar to Principal Component Analysis (PCA)
  - PCA is typically applied to the symmetric covariance matrix of existing variables
  - SVD is usually applied to the relatively sparse document-by-term frequency matrix
- Dimensions of meaning
  - The linear combinations may clearly identify underlying or "latent" dimensions
  - Each term receives a weight denoting its influence on the definition of the respective dimension
  - Mapping documents into dimensions of "meaning" while maintaining the information necessary to differentiate between documents, aka, *Latent Semantic Indexing*.

What do you get?



From: *Taming Text with the SVD*

Example

Original Matrix

	document	error	invalid	message	file	format	unable	to	open	using	path	variable
1	d1	1	1	1	1	1	0	0	0	0	0	0
2	d2	1	0	2	1	0	1	1	1	1	1	0
3	d3	1	0	0	0	1	1	1	0	0	0	1

Weights

	U_2	
error	43	.30
invalid	.11	.13
message	.55	-.37
file	.33	-.12
format	.21	.55
unable	.31	.18
to	.31	.18
open	.22	-.25
using	.22	-.25
path	.22	-.25
variable	.09	.42

Form Dot Products

	document	SVD1	SVD2
1	d1	1.63	.49
2	d2	3.14	-.96
3	d3	1.35	1.64

Normalize

	document	SVD1	SVD2
1	d1	.96	.29
2	d2	.96	-.29
3	d3	.64	.77

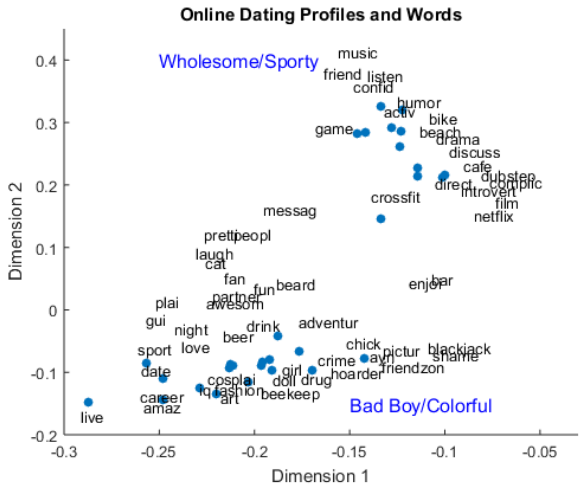
Matrix for Mining

From: *Taming Text with the SVD*

Interpreting Latent Dimensions

- Review the scatter plot of terms against the latent dimensions to derive subjective labels
  - The coefficients for each term along different dimensions
  - Find terms at one or the other end of a dimension
  - Find clusters of terms in the scatter plot
- Validate the dimensions by relating them to other available variables with known meaning
  - Compute the document scores for each document in the latent dimensions
  - Find correlation between the dimensions to other structured information (e.g. product reviews by shoppers published on websites may provide explicit "recommend vs do not recommend" ratings)

Example Plot



From: *Can You Find Love through Text Analytics https://blogs.mathworks.com/loren/2015/04/08/can-you-find-love-through-text-analytics/*

SVD – How Many Dimensions?

- Usually no more than 5 to 20 dimensions extract most of the information from the TDM.
- More dimensions (up to a few hundred) can be retained if the processed data is for subsequent predictive modeling or clustering

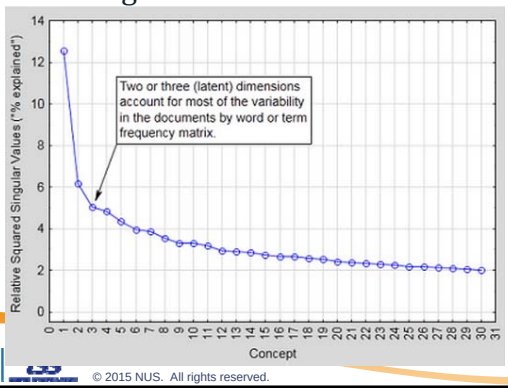


Figure 11.3 Plot of relative squared singular values by number of latent semantic dimensions
From *Practical Text Mining and Statistical Analysis for Non-structured Text data*

Usefulness of SVD

- Good for identifying the latent dimensions of meaning that organize the documents in the corpus
- Generally appropriate for data reduction in text mining
- Not useful if the purpose of the analytical project is to identify the specific phrases or terms that are important and related to key performance indicators (e.g., which phrases in physicians' notes are predictive of subsequent health care costs)
- Computationally expensive

Labeling the Clusters

- Clusters resulting from clustering can be labeled with numbers, which is not very insightful.
- For some applications, for example, browsing, a good, meaningful cluster label is almost as important as good clustering
- Human experts can read and review the assigned labels to understand the results of the clustering process and to reach some decisions about their value.

Labeling the Clusters

- A cluster can be labeled with a small number of carefully selected words distinguishing the cluster from others.
  - Documents are composed of words and the distribution of words is the basis of document clustering
  - We can select:
    - Most frequent words in a cluster
    - Words with largest average *tf-idf* value
    - Using feature selection methods, etc.
- One or more exemplar documents may also be selected as "typical" documents to represent the cluster
  - E.g. the document that is most similar to the cluster mean vector

Evaluation of Clustering Result

- Clustering starts with unlabeled data, therefore the evaluation approach used in classification doesn't apply here. There is no right answer.
- But we can evaluate whether clustering has put similar documents in the same group.
  - E.g. by computing a cluster mean and its variance or standard deviation(error)
  - If documents within a cluster are similar, the **variance or standard deviation of the mean** will be low.

Evaluation of Clustering Result

- For a cluster c with n documents and its mean M_c

$$Variance = \sum_{i=1}^{n} \frac{(D_i - M_c)^2}{n} \qquad SD = \sqrt{Variance}$$

D_i - M_c can be computed using the distance functions discussed earlier.

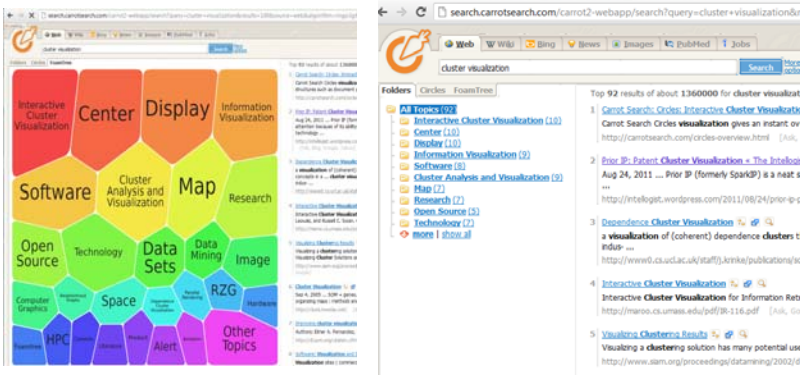
- Another simple way to evaluate in-cluster similarity is to use **shared word counts**.
  - A cluster mean would indicate the average number of words shared between documents in the cluster.
  - Error could be measured by the expected deviation from this number
  - The baseline error would be the one obtained by assigning all documents to the same single cluster.

Intra-cluster Variance vs. No. of Clusters

- With a large number of clusters, each containing few documents, the variance computed for each cluster will be low, but the usefulness of the cluster may be poor.
- How many clusters should the clustering result contain?
  - The higher the number of clusters, the lower the intra-cluster variance.
  - The number of clusters should stop rising when it leads to very small decrease in intra-cluster variance.
  - Or an upper bound of acceptable variance/error can be declared.
  - Or we can also rely on human judgment, which can be subjective.

Document Cluster Visualization

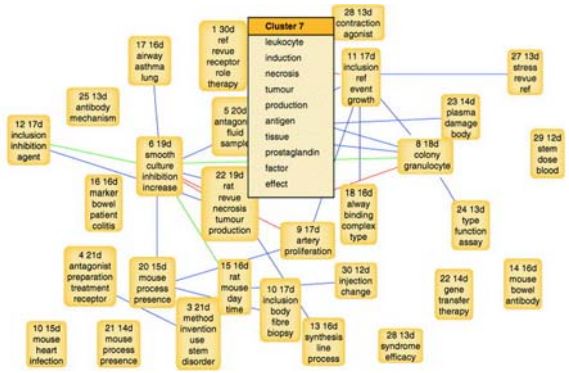
- Visualization can be used to explore the clusters for understanding and evaluation of the clusters.
- Very useful in navigating the clustered search result



Visualization from Lingo3G: text document clustering engine http://carrotsearch.com/lingo3g

Document cluster visualization: another example

- Cluster as node, linked with lines, the strength of which indicating the strength of the semantic relationship of the clusters
- Keywords as descriptions of the cluster topic



Document cluster visualization from Mack & Hehenberger, 2002

Word Clustering

Word Clustering

- Words can be clustered in two ways.

1. By meaning
   - Grouping together semantically similar words into a cluster (or concept)

2. By co-occurrence
   - Grouping words that commonly appear together

Clustering semantically similar words

- It's also referred to as *Concept Extraction* in some literature.
- Useful in grouping and typing domain concepts.
- The context-dependent nature of word meaning

  "You shall know a word by the company it keeps."

  – J. R. Firth (1957)

- Words with similar meaning appear in similar context

  E.g. "dogs", "cats", "fish", "birds", "hamsters"…
  - Referring to household pets
  - Used in the same context

How to cluster semantically similar words?

- Clustering on the similarity between the contexts in which the words appear
- Representing context as a vector space – *term-context matrix*
  - Each term with its context across all documents in the corpus
  - Context – a window of size *n* around the given word (*n-grams*)
  - Representation of context - the count of the words appearing in n-grams
  - N-grams can be counted across a corpus and enhanced by other resources like Google's n-gram datasets from Google Books project
    - http://books.google.com/ngrams/datasets
- Apply clustering algorithms (e.g. *k-means*) with an appropriate distance measure (e.g. *cosine distance*)

New, Alternative Way

- **Word Embedding** – new way to find semantically similar words!
- Learn vector representation of words from lots and lots of text data.
- More details in the next lesson ----

  Stay tuned!

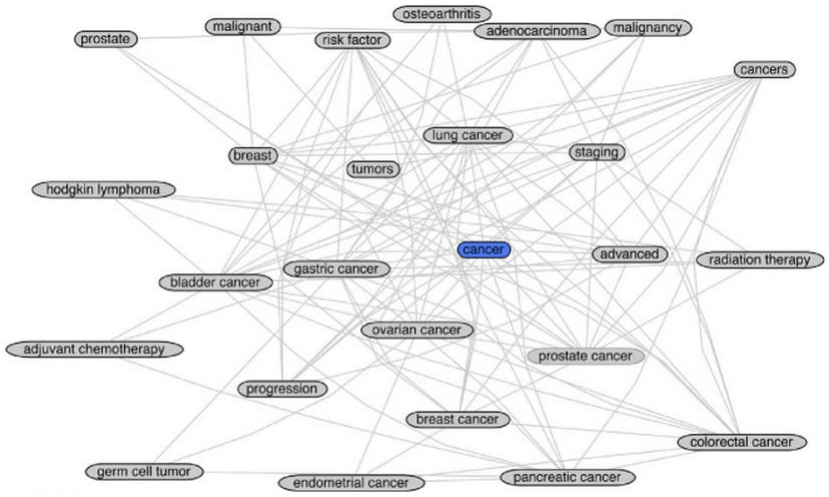
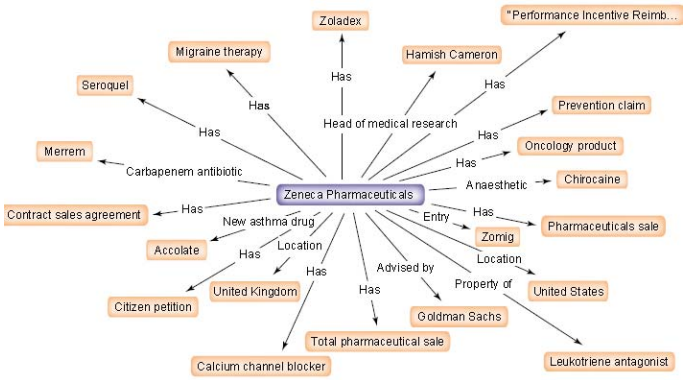


FIGURE 13.1
An example concept inferred from PubMed abstracts using a combination of concept extraction and collocation detection.

From *Practical Text Mining and Statistical Analysis for Non-structured Text data*

Word Cluster Visualization – Similarity Based

- Terms related to the root keyword
- The relationships can be named or unnamed



Lexical Network from Mack & Hehenberger, 2002

Summary

- Clustering is an important technique for data exploration and understanding.
- Cluster requires functions to measure the similarity between data objects, and algorithms to efficiently compare and cluster them.
- Text clustering is used to group together documents or words based on similarity. Document clustering is useful for exploring and understanding how documents are related, whereas word clustering can discover words sharing topical or semantic meanings.

References

- P. Arabie, L.J. Hubert, G. De Soete. Singapore; River Edge (Ed). *Clustering and classification* , NJ : World Scientific, 1996.
- Gary Miner, John Elder IV et. al. Chapter 13 Clustering Words and Documents, *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*, Academic Press, 2012
- Weiss, Indurkhya, & Zhang. Chapter 5 Finding Structure in a Document Collection, *Fundamentals of Predictive Text Mining*, Springer, 2010.
- R. Mack, M Hehenberger. Text-based knowledge discovery: Search and mining of life-sciences documents. *Drug Discovery Today*, 7 (11), 2002
- Albright, Russell. Taming Text with the SVD. SAS, *January* 7, 2004.
- Manning, Chris, and Hinrich Schütze. Collocations. *Foundations of statistical natural language processing* (1999): 141-77.