

WOLFGANG ZINIEL | KARL LEDERMÜLLER | JUNE 25, 2010 | EXPERTENFORUM.AT
**DERIVING CUSTOMERS' PRODUCT PERCEPTION
SPACES FROM FORUM POSTINGS**

Outline

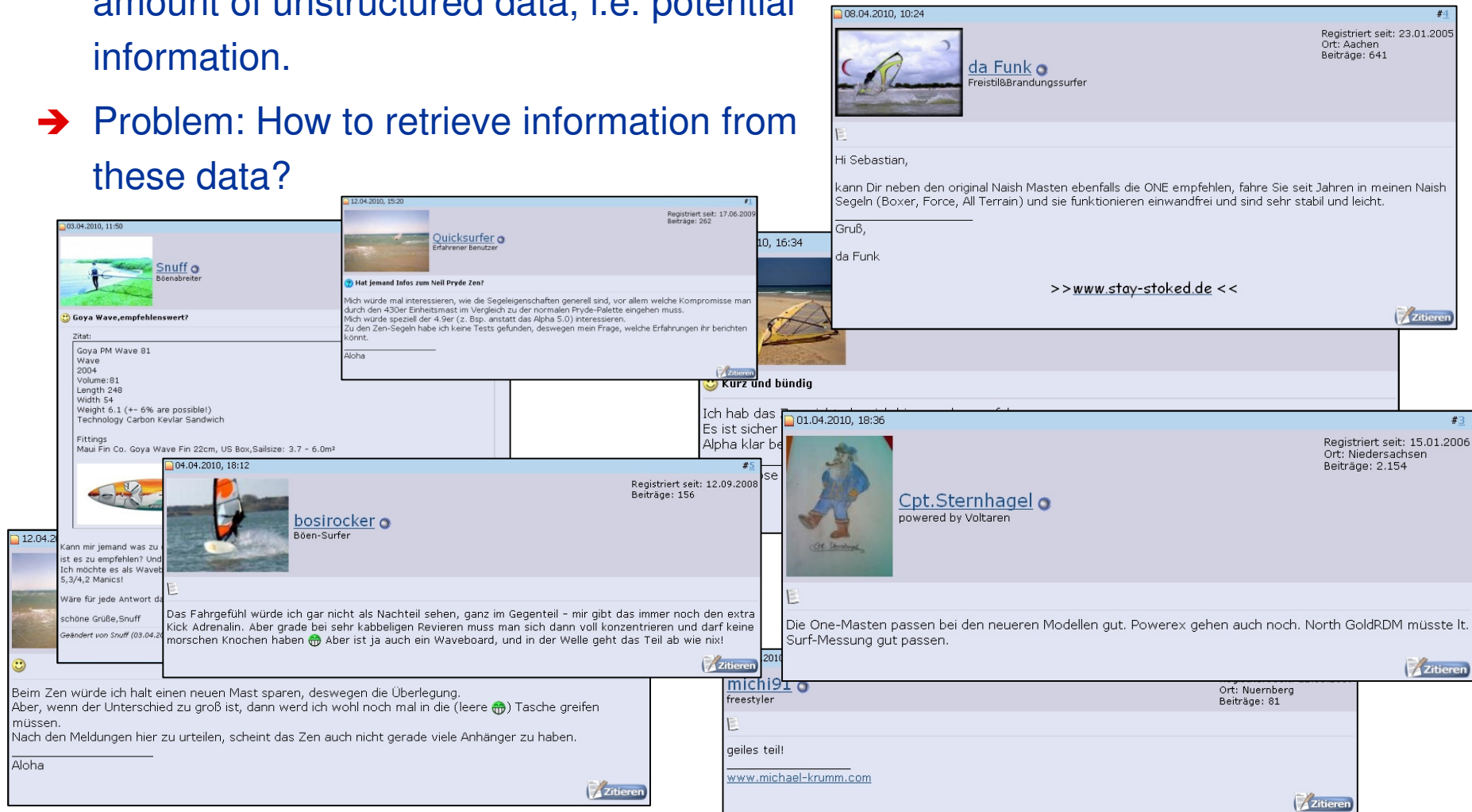
- Research Background
- Aim and Problem
- (Web) Text Mining
- Course of Investigation
- Multidimensional Scaling
- Results
- Issues | Concerns | Discussion

Research Background

- One central endeavor of marketing science is the measurement of product positioning as it is perceived by the customers, i.e. the relative position of a product relative to the competitors
- Regularly data collection is approached in a reactive way, e.g. interviews.
- Mass usage turned the internet into a new social space -> human behaviour can be observed online, too.
- Huge quantity of text in machine readable format is accessible via the internet, e.g. scientific articles, books, mailing lists, blogs or forum postings (Feinerer, Hornik and Meyer 2008).

Aim and Problem

- Initial point: Forum entries represent a huge amount of unstructured data, i.e. potential information.
- Problem: How to retrieve information from these data?



The collage shows several forum posts with the following details:

- Post 1 (Snuff):** 03.04.2010, 11:50. Title: Goya Wave, empfehlenswert? Content: Zitat: Goya PM Wave 81, Wave 2004, Volume: 31, Length: 248, Width: 54, Weight: 6.1 (+/- 6% are possible), Technology: Carbon Kevlar Sandwich, Fittings: Maui Fin Co. Goya Wave Fin 22cm, US Box, Sailsize: 3.7 - 6.0m².
- Post 2 (Quicksurfer):** 12.04.2010, 15:20. Title: Hat jemand Infos zum Neil Pryde Zen? Content: Mich würde mal interessieren, wie die Segeligenschaften generell sind, vor allem welche Kompromisse man durch den 430er Einheitsmast im Vergleich zu der normalen Pryde-Palette eingehen muss.
- Post 3 (da Funk):** 08.04.2010, 10:24. Content: Hi Sebastian, kann Dir neben den original Naish Masten ebenfalls die ONE empfehlen, fahre Sie seit Jahren in meinen Naish Segeln (Boxer, Force, All Terrain) und sie funktionieren einwandfrei und sind sehr stabil und leicht.
- Post 4 (bosirocker):** 04.04.2010, 18:12. Content: Das Fahrgefühl würde ich gar nicht als Nachteil sehen, ganz im Gegenteil - mir gibt das immer noch den extra Kick Adrenalin. Aber grade bei sehr kabbeligen Revieren muss man sich dann voll konzentrieren und darf keine morschen Knochen haben.
- Post 5 (Cpt. Sternhagel):** 01.04.2010, 18:36. Content: Die One-Masten passen bei den neueren Modellen gut. Powerex gehen auch noch. North GoldRDM müsste It. Surf-Messung gut passen.
- Post 6 (michi9):** Content: geiles teil! www.michael-krumm.com

(Web) Text Mining


→ ambiguous definitions

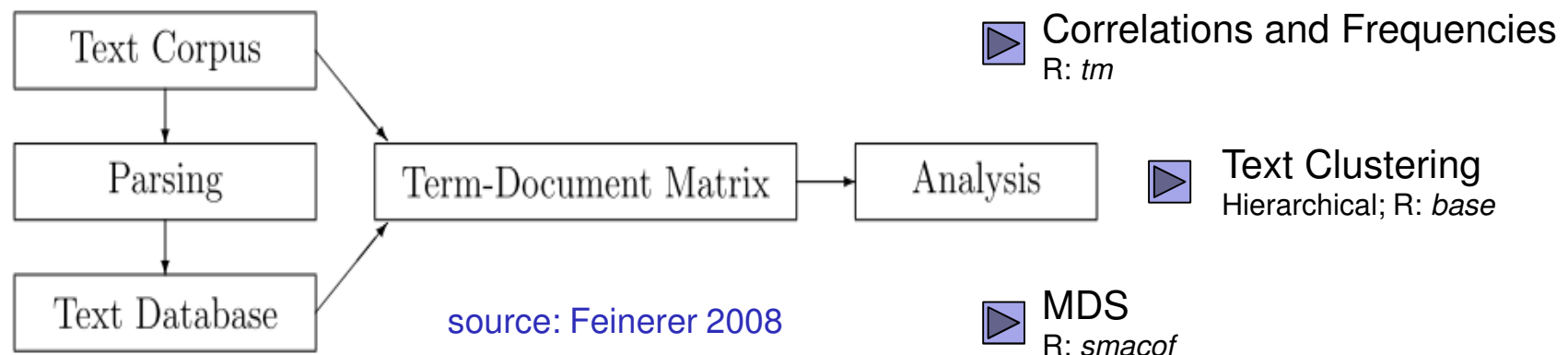
- *“a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools”*
Feldman and Sanger (2007)
- Miller (2005): *“the automated or partially automated processing of text”*
- interdisciplinary technique, including data mining, computational statistics, linguistics and computer science (Feinerer, Hornik and Meyer 2008)

→ textual information from the www -> web text mining

- web usage mining
- web structure mining
- web content mining = *“extracting useful information from unstructured web texts”* (Weiss 2005)

Statistical Methods

- Executable within : packages **base** (<http://cran.r-project.org/>), **tm** (Feinerer 2010) and **smacof** (de Leeuw, Mair 2009)
- **Text clustering**
 - K-means or hierarchical clustering (Willett 1988), based on a transformation of the text into a structured format, e.g. term-document matrix
 - Grouping news articles or information service documents
- **Text classification / categorization**
 - K-nearest neighbor classification, Bayes classifiers or decision trees
- **Information Extraction and Retrieval**
 - Extracting information out of texts, a technique commonly used by modern internet search engines



Course of Investigation 1/2

- ➔ crawling and locally depositing relevant forum entries (by awk and sed)
 - <http://forum.surf-magazin.de/> forum of the German Windsurf magazine
 - 270.000 contributions, important online network for windsurfing
 - 1.785 contributions related to quality issues and sails were chosen
- ➔ creating a corpus (text repository) for the 1785 contributions
- ➔ pre-processing tasks
 - strip whitespace () -> ()
 - remove punctuation (. ; : ,) -> ()
 - remove numbers (123456) -> ()
 - convert upper case to lower case (TEXT) -> (text)
 - remove predefined German stop words (der die das und oder) ->
 - <http://snowball.tartarus.org/algorithms/german/stop.txt>
 - remove standard elements that occur in every post
 - "gruss", "grüße", "icq", "nachricht", "beitrag", "beiträge", "standard", "schicken", "anzeigen", "ort", "zitat ...

Course of Investigation 2/2

- Porter stemming for German text sources
 - <http://snowball.tartarus.org/algorithms/german/stemmer.html>
 - reduces word information to radical by deleting suffixes and by removing umlaut accent (üäö) -> (uao)
 - kategorien -> kategori
 - kategorisch -> kategor
 - katzensprung -> katzenspr
- document-text-matrix
 - initial point for statistical analyses
 - clustering
 - finding associations
 - multidimensional scaling
 - word frequencies

Multidimensional Scaling (MDS)

→ is able to

- identify key dimensions that trigger respondents' evaluations of objects,
- site objects in a perceptual space on the basis of the objects' perceived similarities (also: similarity structure analysis, perceptual mapping),
- determine the perceived relative image of the objects and
- transform customer judgments of overall similarity or preferences into distances in multidimensional space.

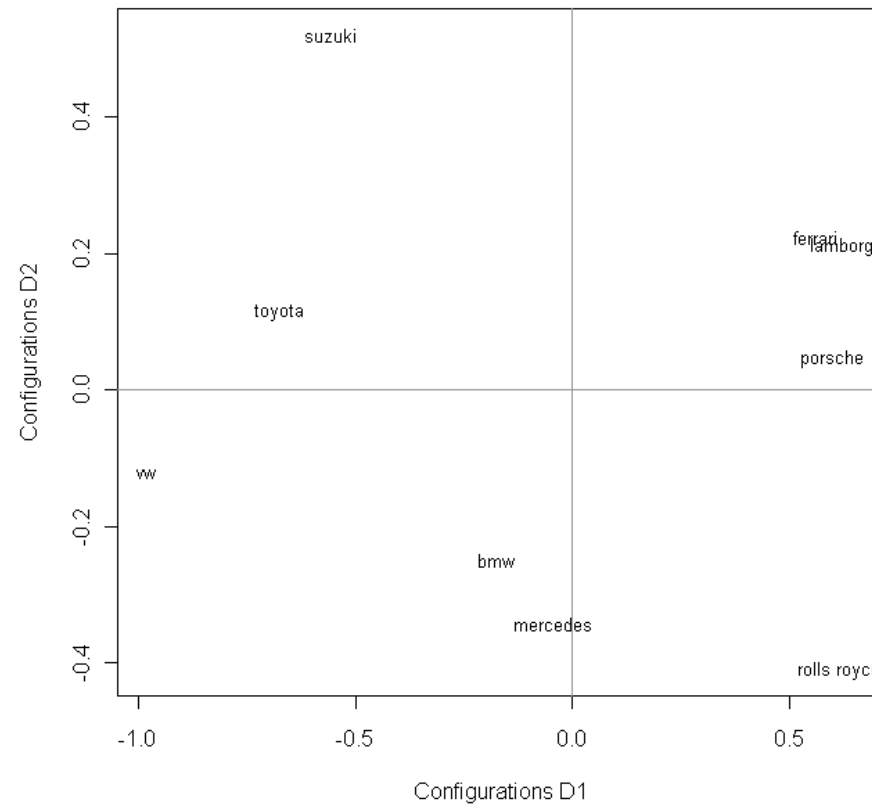
MDS Example

	vw	suzuki	toyota	mercedes	bmw	ferrari	porsche	lamborg.	rollsroyce
vw	0								
suzuki	4.4	0							
toyota	3.3	3.7	0						
mercedes	5.8	7	5.3	0					
bmw	5.5	7	4.2	2.7	0				
ferrari	8.1	8.3	8.3	6.9	6.8	0			
porsche	8.1	8.4	8.3	6.4	6.4	3	0		
lamborghini	8.2	8.8	8.7	6.6	6.4	2.1	3.4	0	
rollsroyce	8.6	8.9	8.2	5.8	7	6.6	6.8	6.3	0

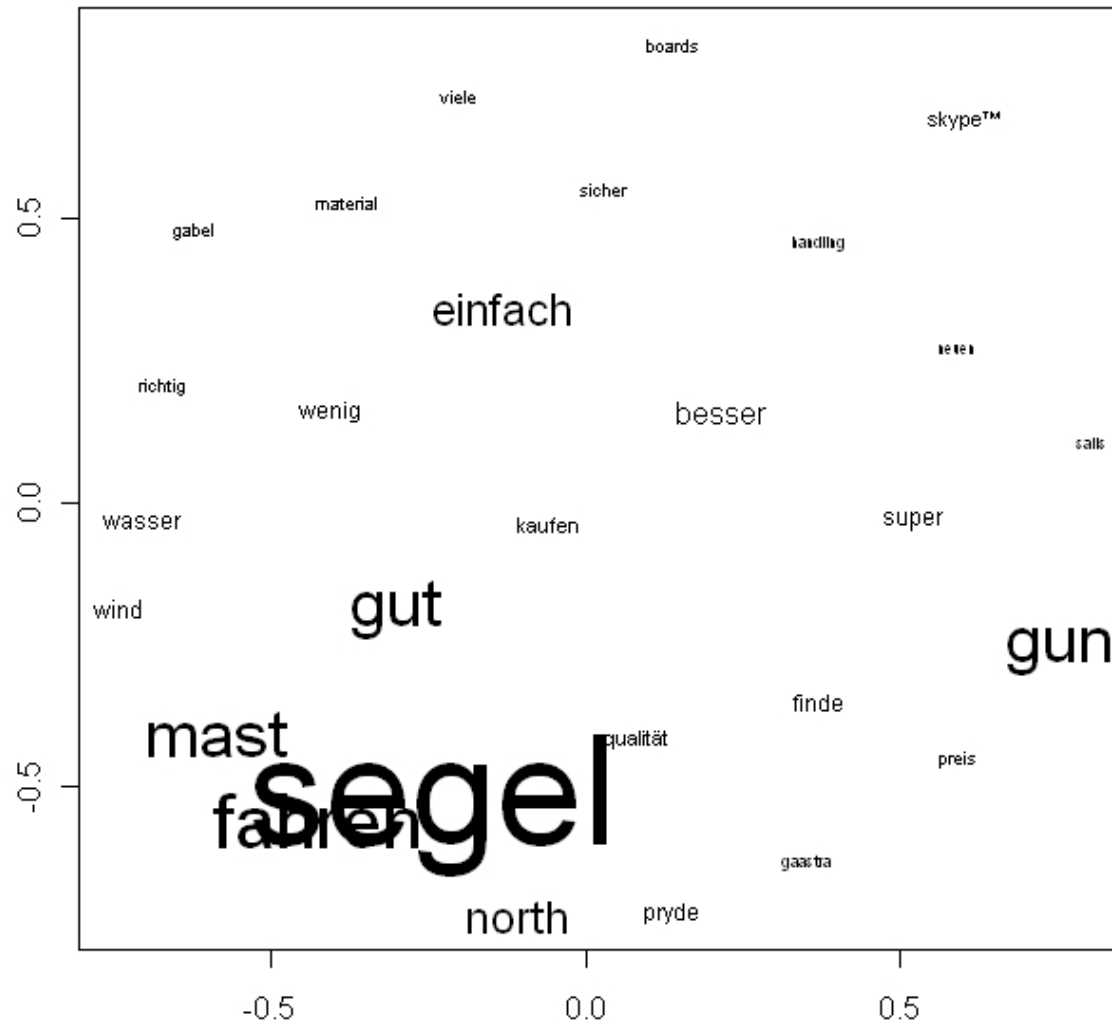
Configuration Plot

“Please rate the similarity of the following car brands using the numbers 1 (= very similar) to 9 (= very dissimilar).”

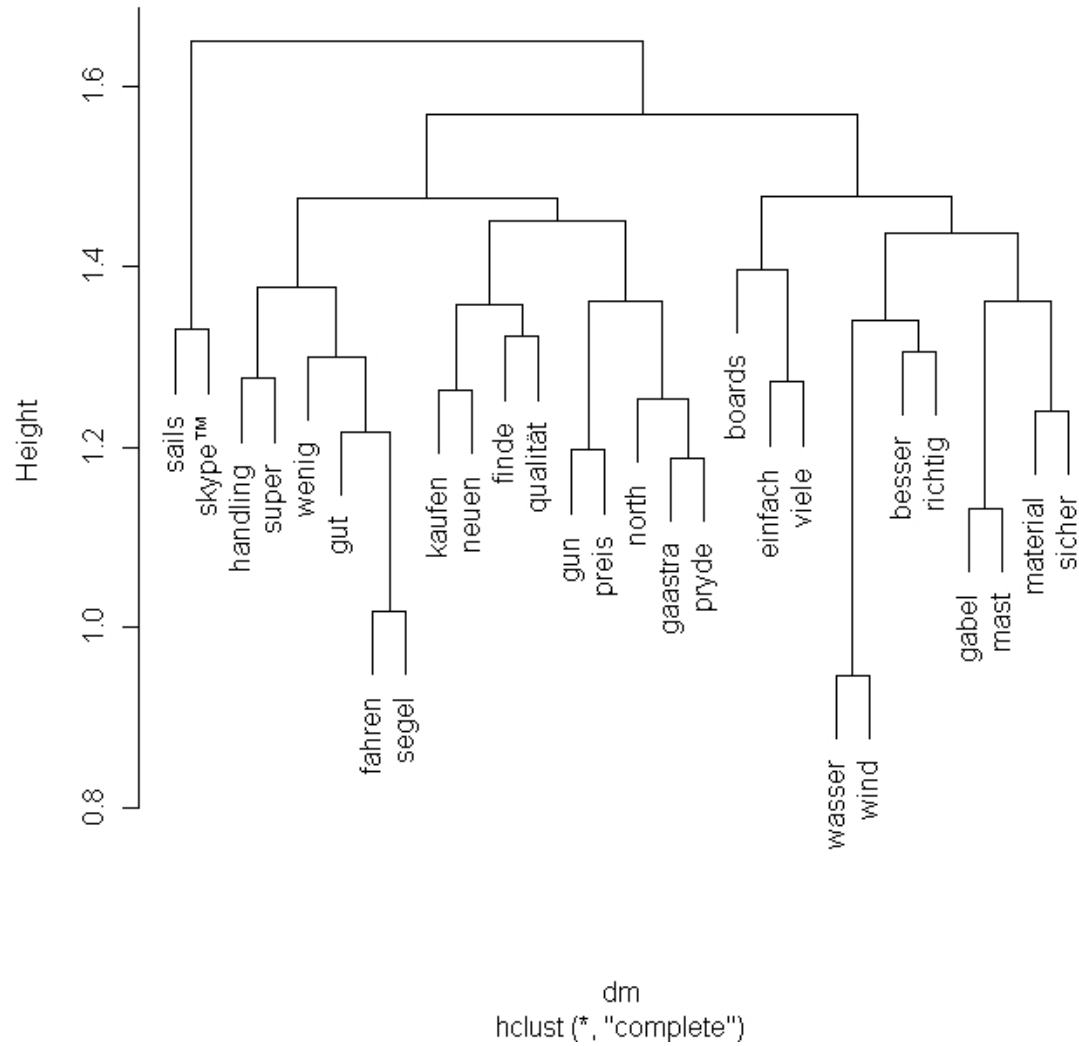
metric stress: 0.0063



RESULTS



Cluster Dendrogram



Issues | Concerns | Discussion

- Broaden the applicability of Text Mining methods within the field of marketing science -> deeper understanding of the consumer.
- Employ quantitative methodology to qualitative problems.
- There is still a need to improve pre-processing procedures; i.e. considering negative phrasing
- *tm* offers a valuable toolkit for broad Text Mining problems.

Bibliography

- Feinerer, I.** (2010). tm: Text Mining Package. R package version 0.5-3, <http://cran.r-project.org/web/packages/tm/index.html> .
- Feinerer, I., Hornik, K., Mayer D.** (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25(5): 1-54.
- Feldman, R., Sanger, J.** (2007): The Text Mining Handbook. Cambridge University Press.
- Hearst, M.** (1999). Untangling Text Data Mining. 37th annual meeting of the Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics.
- de Leeuw, J., Mair, P.** (2008): Multidimensional Scaling Using Majorization: SMACOF in R. UCLA Working Paper Series, <http://repositories.cdlib.org/uclastat/papers/2008010903> and <http://cran.r-project.org/web/packages/smacof/> .
- de Leeuw, J., Mair, P.** (2009). Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software, 31(3), 1-30. URL <http://www.jstatsoft.org/v31/i03/> .
- Miller, T. W.** (2005). Data and Text Mining.
- Weiss, S., Indurkha N., Zhang, T.** (2005): Text mining - Predictive Methods for Analyzing Unstructured Information. Springer.
- Willett, P.** (1988): Recent trends in hierarchic document clustering: a critical review. Information Processing and Management, 24(5): 577-597.