Tuning Text Classification for Hereditary Diseases with Section Weighting

Jörg Hakenberg [a] **Juliane Rutsch** [b] **Ulf Leser** [a]

[a] Knowledge Management in Bioinformatics, Humboldt-Universität zu Berlin,

[b] School of Electrical Engineering and Computer Science, FH Stralsund

Abstract

Motivation: *Information in life science publications is heterogeneously distributed over various sections. Depending on research questions, different sections cover more or less of the data needed to answer them. Our approach, called section weighting, seeks to make use of information coverage and density found in typical life science publications. We study the impact of section weighting on text classification according to hereditary diseases.*

Results: *Our results indicate that weighting sections can improve text classification. Our systems gain 7% in F1-measure when we add section weighting. Proper composition of features is equally crucial, improving our results by 11%. Combining both techniques, the system yields a performance 18% higher than the baseline classifier. For our research question, favoring the sections Abstract, Introduction, and Materials and Methods yields the best results.*

Introduction

Analysis of biological and clinical research publications as a major knowledge repository is of increasing importance. The large amount of publications in this fields necessitates automated systems to aid researchers. MEDLINE, a well-known citation index, currently contains almost 16 million references. Despite growing numbers of results stored in biomedical data bases, still only the publication of research findings in journals is deemed reputable and prestigious. Data bases provide fast, well-defined, and easy access to structured information. Texts are semi-structured only, and problems arise when automated tools parse such data. Manual browsing through large amounts of textual information, on the other hand, is an infeasible task. Keeping up with recent and relevant developments gets more and more complex for researchers in the life sciences. At first there is only the possibility for full text searches (*e.g.* as provided by PUBMED) to obtain relevant publications. Searches for exact keyword matches often fail in terms of recall and precision. Different usage of nomenclatures or even simple spelling variants lead to serious information-losses. Homonyms and ambiguous words referring to irrelevant concepts are found during keyword searches nevertheless, resulting in large result sets containing irrelevant documents. These problems render manual searches in large citations indexes infeasible, and any automated approaches become potentially significant for improvements in terms of time and costs.

Text classification is the task of assigning one or more categories to arbitrary texts, based on models learned from labeled examples (*supervised learning*). Labels function as short descriptions for texts, and users can easily look into all texts annotated with relevant labels, instead of searching for texts themselves. In addition to searching for relevant publications, we see text classification as a filter component and building block for further text processing and information extraction. Whenever researchers try to find genes associated with a particular disease, it is infeasible to parse all publications for gene names, for instance. Intuitively, one would first reduce the text collection to all publications discussing this particular disease. Subsequent *information extraction* such as *named entity recognition* (NER) or *relation mining* can then be applied to the much smaller remainder of relevant texts.

The task of classifying texts to filter publications discussing hereditary diseases does not conclude after finding *any* relevant publications, but includes the categorization into specific diseases. While binary text classification (relevant/irrelevant) yields very good results, multi-class supervised learning proved much harder. Classifiers applicable to multi-class learning have been proposed over the last years (see, *e.g.*, the WEKA library [Frank *et al.*, 2004]). The introduction of meta-learners, such as boosting or voting schemes, often improves results, but nevertheless performances stall. In this paper, we do not present a new supervised learner, but study improvements of existing methods. In particular, we address weighting of sections and variations of feature generation.

We sought to improve classifiers by weighting typical sections of papers differently, with respect to the information density. For example, an abstract is a concise description of a text, and thus should mention the main concerns of the authors, including names of diseases and treatments. Shah *et al.* (2003) showed that the keyword content of different sections is very heterogenous. Sinclair and Webber (2004) studied text classification using single sections instead of full texts. Our approach used heuristic weighting schemes, empirically gathered weights, and data on information distribution previously published by other groups [Shankar and Karypis, 2000, Schuemie *et al*., 2004].

We studied the generation of features for representing documents using different *natural language processing* techniques (NLP). The *bag-of-words* representation with token counts is altered using *stemming*, *stopword filters*, and *part-of-speech information*. In addition, we removed useless structural information and mapped appearing numbers to generalized tokens.

In this paper we present results for the text classification applied to full texts discussing hereditary diseases. For our evaluation, we took 25 hereditary diseases and their respective descriptions from OMIM. We used 22 publications referenced in these descriptions as training documents for these classes. We parsed all texts for the seven most common sections, and changed the default weighting scheme of tokens (token counts, weighted with $tf{\cdot}idf$) according to relevance of sections. For the vocabulary we used either all tokens, only nouns, only nouns and verbs, or only their respective stems.

Related Work

Some approaches for multi-class learning problems and applications to text classification have been studied, see *e.g.*, the work of [Joachims, 1998, Han and Karypis, 2000, Raychaudhuri *et al*., 2002]. We discuss an approach classifying texts using only individual sections, and work presenting typical information distributions in biomedical texts.

Sinclair and Webber (2004) tried to automatically associate articles with GeneOntology (GO) codes. They studied which sections were most relevant, and which NLP techniques were most useful. Starting with HTML formatted and well-structured articles from BMC, markup of sections was done by hand. The *Naïve Bayes* classifier was trained on Title and Abstract of the texts, split into nine classes (GO codes), and tested on individual sections or full texts, respectively. The maximum f-measure could be achieved for classifying Titles only (64.7%). The highest recall was 89.8% on full texts with stemmed words. The system yielded its maximum precision of 65.4% on Titles. A conclusion from their work is that the Materials&Methods section was comparably valuable in contrast to Shah's findings. Our own system was trained and tested on complete full texts. Not taking individual sections into account for classification would be similar to setting their respective weights to zero. Our task definition included 26 classes (diseases plus negative examples), and results are not fully comparable. Information distribution concerning GO terms (describing biological processes, cellular components, and molecular functions) surely is different.

Shah *et al.* (2003) showed that keyword content in different sections was very heterogenous. Abstracts provided the highest keyword density, but other sections might be better sources for the extraction of biologically relevant data. There results were based on a set of 104 full texts. They selected keywords basically using correlation analysis of words within a section. Shah *et al.* conclude that for particular tasks some sections should be avoided. Methods proved to be the worst place to search for gene and protein names, for instance. For mining biological concepts (species, tissues, diseases, etc.), they preferred Discussion and Results over Abstract and Introduction. Concerning technical data, chemicals, and measurements, Methods sections were most appropriate.

Schuemie *et al.* (2004) analyzed 3902 full text articles manually for information densities and coverages. They studied information distribution of different semantic types (chemicals & drugs, genes, diseases, organisms) over different types of sections. They conclude that the popular restriction to only Abstracts for information retrieval and extraction leads to serious information-loss. Their results show that the information density is highest in Abstracts, and lowest in Methods sections. Coverage was highest in Results sections (30-40%). Concerning chemicals and drugs, the Methods sections were richest in information, while diseases and genes were mentioned less frequently here.

This paper is organized as follows. We first present text classifiers we studied, feature generation, methods for detecting the current section context, and our evaluation data set. Afterwards, we show the impact of weighting, classifiers, and feature generation and other results. This paper concludes with a discussion and presentation to related work.

Approach

We represented each document in a vector space constructed of all features appearing throughout the document collection. All features occurring in the training corpus were stored in a *feature vector*, *i.e.* representing the *bag-of-words vocabulary*.

The *document vector* for each text was constructed as follows. Reading a given document, we generated features from the tokens we encountered (see below). If a feature appeared in the feature vector (*i.e.*, it occurred in the training corpus), we added it to the current document's vector, and in the following counted this features occurrences. In an unweighted scheme, each oc-

Table 1: Word counts of distinctive keywords relevant to *Retinoblastoma* in proportion to the number of all words for each section. The table shows distributions in one document and averaged over 20 documents. RB, abbreviation for Retinoblastoma; RB1, gene/protein name; pRB, retinoblastoma protein; 13q14, gene locus.

	Abs	Intro	M&M	Results	..
Retinoblastoma	2/359	1/439	1/1004	0/919	
RB	1/359	2/439	1/1004	1/919	
RB1	3/359	11/439	3/1004	10/919	
pRB	4/359	4/439	4/1004	9/919	
13q14	0/359	1/439	1/1004	0/919	
..					
Avg. (13 words)	.0077	.0096	.0027	.0076	
Avg. (20 docs)	.0143	.0097	.0044	.0105	

currence in a text counts one, with the final counts for every feature representing the document. In our approach, we weighted each feature using *tf·idf* (term frequency, *tf*, corresponds to our feature counts).

To alter the weight of different sections, we increased the count not by one when encountering the corresponding feature, but used different addends/factors for each section. When a feature appeared in one section, it added more (or less) to the feature count than its appearance in another section did. The final weighting with *tf·idf* remained the same.

In our approach, we used heuristic weighting schemes, empirically gathered weights, and data on information distribution previously published by other groups [Shankar and Karypis, 2000, Shah *et al.*, 2003, Schuemie *et al.*, 2004, Sinclair and Webber, 2004]. Heuristic weights were based on the intuitive information content for different sections. We computed weights empirically by looking into the information distribution concerning relevant keywords (names of diseases, genes, typical symptoms, etc.). We collected such data for three classes, with up to 15 keywords each. Table 1 shows examples for the disease Retinoblastoma.

Section Context

The weighting of features depended on the text section where they occurred. Accessing each document word-after-word, it was necessary to detect section headings and recognize changes of the current context. Research publications most often follow a similar structure, though ordering might differ. They start with titles and authors' names and affiliations, followed by a short abstract, an introduction to the topic and so on. The main structural building blocks found in (biomedical) scientific publications are *Opening*[1], *Abstract, Introduction, Materials and Methods, Results, Discussion, and References*. Problems occurred not with the ordering of sections, but the proper identification of (relevant) section headings. While it was easy to solve variations in capitalization, detecting alternative names (*e.g.*, 'References' vs. 'Literature') for sections was only possible after manual inspection of publications. Table 2 lists examples for variants of section headings. Another problem occurred in HTML formatted documents. It was not sufficient to switch the context after each occurrence of a heading corresponding to a section name. Hyperlinks in an HTML document pointing to other sections within the same document often get the same name as the section they link to.

Table 2: Examples for varieties and spelling variants for sections typically labeled 'Materials and Methods' and 'References'.

Materials and Methods	References
Materials and methods	REFERENCES
material and methods	references
Patients and Methods	References & Notes
Patients and methods	LITERATURE
PATIENTS AND METHODS	
METHODS	
Subjects and Methods	

Feature Generation

Orthogonal to the weighting of features according to their occurrence in the texts, we evaluated different methods of feature generation. The simplest and most common form of features for text classification are tokens appearing in a text collection. Most often, their weights are calculated using token counts or *tf·idf* weighting. We studied the influence of different processing and filtering steps on classification performance. *Tokenization* depends on word boundaries, for which we took '.', ',', ';', ':', '?', '!', brackets, blanks, and newlines. *Word stemming* of tokens is done using an own implementation of the Porter stemming algorithm [Porter, 1980]. We filtered different amounts of *stop words*, either 100, 1000, or 10.000 common English words [wortschatz lexikon, 2004, Biemann *et al.*, 2004]. In order to reduce the set of tokens to all nouns, or all nouns and verbs, we use QTAG [Tufis and Mason, 1998] for *part-of-speech tagging*. We mapped all numbers and percentages to generalized tokens, <NUMBER> and <PERCENTAGE>, respectively. All documents in the data collection were

[1] Title, Authors, Affiliation, Journal, etc.

stored in HTML format, so that HTML tags and other non-informative text blocks had to be filtered. Such text blocks were, *e.g.* links for navigation within the page, which often had the same name than the section (heading) they link to.

Text Classifier

For our experiments, we applied a *Nearest-Centroid* classifier (NCC) to the multi-class problem on our data set. Han and Karypis (2000) showed high performances for centroid-based classifiers, based on 20 different data sets (Reuters, OshuMed, Trec, etc.).
The predicted label (*i.e.* class) of nearest-centroid classification is the label of the centroid lying next to the new example in the vector space. *Centroids* are mean vectors for each class, calculated as the average from all representatives for this class:

$$\forall c_i \in C : z_i = \frac{\sum_{j=1}^{N} d_j}{N},$$

where C is the set of classes, N refers the number of documents d_j in a class c_i, and z_i depicts the centroid of class i. As distance measure between two documents (*i.e.*, a new example and a centroid) we took the cosine distance,

$$dist(x,y) = 1 - \frac{x \circ y}{\|x\| \cdot \|y\|},$$

between two vectors x and y, using their dot-product ($\circ$) and norms ($\|x\|$). The predicted label of a new document d' is the label of the closest centroid found among all class centroids:

$$h(d') = \operatorname*{argmin}_{z_i \in Z} dist(d', z_i)$$

Data Set and Evaluation

The documents we chose for evaluating our approach were taken from the Omim-Database, an online catalog for human genes and genetic disorders [OMIM, 2000]. To obtain an evaluation corpus, we collected 22 documents for each of 25 different hereditary diseases (listed in Table 5). We started with the base entry (*i.e.* a general description) for each disease in Omim, and found other relevant documents using publications cited in these base entries. We added 21 citations for which we were able to obtain full texts (*e.g.*, following links from Omim to PubMed and finally to the publisher's site). Because PDF to ASCII plain text conversion sometimes is erroneous, we looked for the first 21 documents available in HTML format. We added a class consisting of 33 negative examples (documents mentioning none of the 25 diseases). We provide the document collection at http://www.informatik.hu-berlin.de/~hakenber/publ/suppl/. We performed a 10-fold cross-validation, training on $\frac{2}{3}$ of the collection and evaluating on the remainder in each iteration.

Table 3: Results for combinations of individual section weights. Column[1] corresponds to the example in Table 1.

Opening	1	0.3[1]	0.5	1	1	1	0.4	1
Abstract	1	1.4	5	5	3	3	0.6	3
Intro	1	1.0	5	5	3	3	0.9	3
Mat&M.	1	0.4	3	5	1	1	0.6	1
Results	1	1.0	1.5	1	1	1	0.6	1
Discussion	1	1.6	1.5	1	3	1	1.0	1
References	1	0.3	1.5	1	1	3	0.5	1
F1 in %	67	68	68	70	71	71	71	74

Implementation All methods for preprocessing and feature generation as well as classifiers were implemented in Java, the class library being available on request. On a standard workstation with 2.8GHz and 2GB main memory, each 10-fold cross-validation run (train and test) lasted from 10 to 45 minutes, with an even split between preprocessing and learning.

Results

As baseline for all experiments we chose nearest-centroid classification with simple bag-of-words representation of documents. We took all tokens as they appeared in the documents, and weighted them with default $tf \cdot idf$. This systems yielded 60% precision at 53% recall (micro-average; F1-measure: 56%). Altering preprocessing, the system obtained 71% precision at 64% recall (F1: 67%). This included filtering of 10.000 stop-words, and reducing the vocabulary to only nouns.
In addition, we used generalized tokens for numbers and percentages. Stemming of nouns, thus ignoring declensions and plurals, did not improve results any further. Adding section weighting, the system yielded 77% precision at 71% recall (F1: 74%). This weighting favored the sections Abstract and Introduction and yielded the highest precision. The highest recall rates (>70%) can be achieved when weighting Abstract, Introduction, and Materials and Methods higher than the other sections. Some other weight distributions performed comparably well, see Table 3.
Ranking the sections Abstract, Introduction, and Materials and Methods higher than the others improved prediction performances in most cases. Table 3 presents the most useful combinations of weights. In Table 4 we show results for altering the weight of only one section, while all others get the same weight.

Discussion

Our results give a clear indication that weighting sections improves text classification with nearest-centroid

Table 4: Results for combinations of weights where all but one section got the same weight.

Opening	1	1	1	1	1	1	1	1
Abstract	1	1	1	1	1	1	1	1
Intro	1	1	1	1	3	1	1	1
Mat&M.	1	2	3	5	1	1	1	1
Results	1	1	1	1	1	3	1	1
Discussion	1	1	1	1	1	1	3	1
References	1	1	1	1	1	1	1	3
F1 in %	67	65	70	65	68	62	66	69

Table 5: Hereditary diseases annotated in OMIM comprising our evaluation corpus (25 text documents each). Recall and precision values (in %) for each single class.

Disease	Rec.	Prec.
Alopecia	71	56
Alport Syndrome	86	67
Asthma	100	64
Ataxia Telangiectasia	86	75
Cat Eye Syndrome	100	64
Charcot-Marie-Tooth Disease	86	75
Cri du Chat	86	60
Deafness	86	75
Gaucher Disease	86	86
Glaucoma	100	78
Macular Dystrophy	57	100
Neurofibromatosis	14	100
Obesity	71	50
Phenylketonuria	71	100
Porphyria Variegata	57	57
Prader-Willi Syndrome	29	100
Retinoblastoma	14	100
Rett Syndrome	60	60
Schizophrenia	71	71
Severe Combined Immundeficiency	86	100
Sickle Cell Anemia	100	78
von Hippel-Lindau Syndrome	86	100
Werner Syndrome	71	71
Wilson Syndrome	29	50
Zellweger Syndrome	86	100
Negative control	59	56
Average (F1=74)	71	77

classifiers. Our system gains 7% in F1-measure when we add section weighting. Proper composition of features is equally crucial, improving our results by 11% (precision, recall, and F1). Combining both techniques, the system yields a performance 18% higher than the baseline classifier.

Information with discriminative power is not distributed equally within texts. However, this method is only useful for text classification, and bears no direct impact for further information extraction. Numbers for information density and coverage are published or can easily be computed, thus adapting weights to the exact task definition. For the classification of publications concerning experimental setting, treatment or therapy, or other topics, weights different from our best setting can be more appropriate. Looking into our own data set, experimental settings are rarely discussed in abstracts or introductions, and treatments often occur only in the results or discussion.

The common view that the Materials and Methods section found in scientific articles most often contains no relevant information is contradictory to our finding. Higher weights for this section (compared to other sections) led to higher prediction performances. We argue that the research articles studied for our approach contain a relatively high percentage of clinical studies. Materials and Methods (quite often referred to as Patients and Methods) thus describe characteristic symptoms, evidences, tests and experiments, or even measurements. This is the reason for the occurrence of comparably much keywords with discriminating power in this section.

We currently look into the representation of whole sections as different parts in the vector space. At the moment, all features (tokens) are mapped to a single dimension. Classifiers applicable to high-dimensional feature spaces, such as *Support Vector Machines*, could then learn weights for features with respect to their occurrences. Computation of weights or trial-and-error approaches needing manual intervention would be reduced to a minimum. A more systematic approach, however, might ultimately yield the set of weights performing best on our and any arbitrary text collections. If the set of weights proves to be highly specific for each data set, search algorithms or other optimization strategies have to be considered to automatically find an optimal solution.

Centroid-based classifiers allow for the introduction of a minimal distance between new examples and centroids. If a new example is further apart from every centroid than this minimal distance, it is not assigned the label of the nearest centroid, but none at all (meaning no disease is assigned to this document). In addition, we could look at the *m* centroids lying within a certain (minimal) range around the new example, and assign multiple labels. Some other classifiers allow for similar approaches. We discovered that some publications did not belong to only one class, because they discussed multiple diseases. On the other hand, some diseases depend on more than one factor, *e.g.* genotypic predisposition and environmental factors (for instance, adiposity). Such factors are discussed in publications for multiple diseases, decreasing their discriminating power.

Predictions of Nearest-Centroid and other classifiers can easily be transferred to rankings for labels (or even probabilities). These classifiers provide a ranking of labels, with the most probable label on top. If the correct label of a document is not the predicted one, then it would be interesting which rank the correct label got. Semi-automated text filters would produce more than one prediction, and users can easily validate the results. With little human intervention, the new document adds to the training sample as a labeled example. Proper identification of sections is crucial to the method presented in this paper. We currently rely on fixed headings (and spelling variants) as marks for each of the seven section types. Categorization of sections using their respective content as input data would certainly be a more flexible technique. Methods like *zone identification* (ZI) have been proposed and evaluated, see e.g. [Mizuta and Collier, 2004, Ruch *et al.*, 2003]. Results from the TREC 2004 genomics track showed that the MESH terms included in MEDLINE citations are the best predictors for a proper GO annotation. Whenever a full paper can be resolved to its corresponding MEDLINE citation, including these terms in the input data certainly improves performance. Problems arise whenever text classification systems supply relevance filters for new documents. This is often the case in real-life scenarios, but MESH terms are not assigned immediately to a new publication when a citation is added to MEDLINE.

OMIM provides almost 1700 descriptions of phenotypes and catalogues more than 10.000 human genes with known sequence. All descriptions are annotated with references to the literature, all as links to PUBMED (if available). OMIM proved to be an easy-to-access source for text collections, reducing manual intervention to a minimum. However, base entries sometimes contain references to publications not directly related to the disease discussed (but to common experimental settings, diagnostic methods, etc.).

Acknowledgements:
This work is supported by the German Federal Ministry of Education and Research (BMBF) under grant contract 0312705B. The Knowledge Management in Bioinformatics Group is a member of the Berlin Center for Genome Based Bioinformatics (BCB).

Address for Correspondence:
Jörg Hakenberg, Humboldt-Universität zu Berlin,
Knowledge Management in Bioinformatics,
Unter den Linden 6, D-10999 Berlin, Germany
hakenberg@informatik.hu-berlin.de

References

[Biemann *et al.*, 2004] Biemann,C., Bordag,S., Heyer,G., Quasthoff,U., Wolff,C. (2004) Language-independent Methods for Compiling Monolingual Lexical Data, In *Proc CicLING and Springer LNCS 2945*, 215-228, Seoul, Korea, Springer Verlag Berlin Heidelberg.

[Frank *et al.*, 2004] Frank,E., Hall,M., Trigg,L., Holmes,G., Witten,I.H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479-2481.

[Han and Karypis, 2000] Han,E.-H., Karypis,G. (2000) Centroid-Based Document Classification: Analysis & Experimental Results, Technical report, University of Minnesota, Department of Computer Science / Army HPC Research Center, Minneapolis, Minnesota, USA.

[Joachims, 1998] Joachims,T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proc ECML*, Springer Verlag.

[Mason, 2004] Mason,O. (2004) QTAG - a probabilistic part-of-speech tagger, last visited: Jan 2005
http://www.english.bham.ac.uk/staff/omason/
software/qtag.html.

[Mizuta and Collier, 2004] Mizuta,Y., Collier,N. (2004) Zone identification in biology articles as a basis for information extraction. In *Proc JNLPBA at COLING2004*, Geneva, Switzerland, 29-35.

[OMIM, 2000] Online Mendelian Inheritance in Man, OMIM^{TM}(2000) McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).

[Porter, 1980] Porter,M.F. (1980) An algorithm for suffix stripping, *Program*, 130-137.

[Raychaudhuri *et al.*, 2002] Raychaudhuri,S., Chang,J.T., Sutphin,P.D., Altman,R.B. (2002) Associating genes with GO Codes using a maxent analysis of biomedical literature, *Genome Research*, **12**, 203-214.

[Ruch *et al.*, 2003] Ruch,P., Chichester,C., Cohen,G., Coray,G., Ehrler,F., Ghorbel,H., Müller,H., Pallotta,V. (2003) Report on the TREC 2003 Experiment: Genomic Track. *TREC 2003*.

[Schuemie *et al.*, 2004] Schuemie,M.J., Weeber,M., Schjivenaars,B.J.A., van Mulligen,E.M., van der Eijk,C.C., Jelier,R., Mons,B., Kors,J.A. (2004) Distribution of information in biomedical abstracts and full-text publications, *Bioinformatics*, **20**, 2597-2604.

[Shah *et al.*, 2003] Shah,P.K., Perez-Iratxeta,C., Bork,P., Andrade,M.A. (2003) Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, **4**.

[Shankar and Karypis, 2000] Shankar,S., Karypis,G. (2000) Weight adjustment schemes for a centroid based classifier, Technical Report, Department of Computer Science, University of Minnesota, Minneapolis, Minnesota, USA.

[Sinclair and Webber, 2004] Sinclair,G., Webber,B. (2004) Classification from Full Text: A Comparison of Canonical Sections of Scientific Papers, In *Proc JNLPBA at COLING2004*, Geneva, Switzerland.

[Tufis and Mason, 1998] Tufis,D., Mason,O. (1998) Tagging Romanian Texts: a Case Study for QTAG, a Language Independent Probabilistic Tagger, In *Proc LREC*, 589-596, Granada, Spain.

[Witten and Frank, 1999] Witten,I.H., Frank,E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 1999.

[wortschatz lexikon, 2004] Abteilung Automatische Sprachverarbeitung, Universität Leipzig, wortschatz lexikon, last visited: Jan 2005.
http://wortschatz.informatik.uni-leipzig.de/index.html.