# Textklassifizierung nach Erbkrankheiten aus OMIM

## Exposé einer Diplomarbeit

betreut von: Prof. Ulf Leser, Jörg Hakenberg

bearbeitet von: Juliane Rutsch

September 2004 – November 2004

## Problemstellung

Medizinische Fachartikel aus einer Online-Datenbank sollen bezüglich ihres Inhaltes klassifiziert werden. Die zur Verfügung stehenden Volltexte beschreiben Erbkrankheiten. Anhand von charakteristischen Merkmalen soll ein Modell gelernt werden, das einzelne Publikationen anhand des Textes einer bestimmten Krankheit zuordnen kann. Für die Analyse der Volltexte werden Verfahren zur Verarbeitung natürlicher Sprache und des maschinelles Lernen [9] benutzt.

Das Hauptproblem dabei ist, geeignete Merkmale zu definieren und zu gewichten, die eine Gruppe von Texten von anderen Gruppen von Texten bezüglich ihres Inhaltes unterscheidet.

## Zielsetzung

Im Rahmen der Diplomarbeit wird ein konkretes Modell entwickelt, das durch maschinelles Lernen lernt, medizinische Fachartikel den entsprechenden Klassen von Erbkrankheiten zuzuordnen. Das Modell soll auch Texte richtig zuzuordnen können, die nicht als Trainingsdaten benutzt wurden.

## Vorgehen

Zunächst wird in der OMIM - Datenbank (OMIM = Online Mendelian Inheritance in Man ) [1] eine Recherche durchgeführt, bei der Volltexte ausgesucht werden, die sich mit Erbkrankheiten beschäftigen. Dabei beschränken wir uns vorerst auf 25 Erbkrankheiten, zu denen jeweils etwa 20 Fachartikel (soweit verfügbar) gesammelt werden. Jede Krankheit bildet eine eigene Kategorie (Klasse). Die 20 Artikel (Dokumente) zu jeder Krankheit werden in etwa 15 Trainingsfälle und 5 Testfälle aufgeteilt (siehe Abbildung 1).

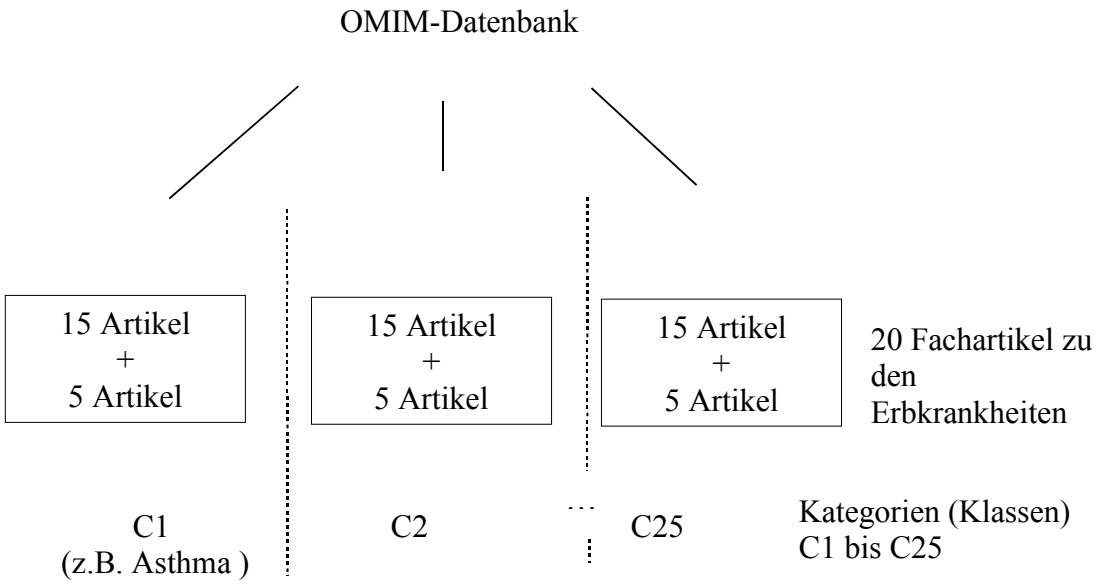


*Abbildung 1: Übersicht über Kategorien*

Ziel ist es, eine Unterscheidung zwischen den Klassen für beliebige Dokumente treffen zu können. Dazu werden die Eigenschaften der einzelnen Dokumente (Trainingsfälle) untersucht.

Die Bestimmung von geeigneten Eigenschaften, die Publikationen über eine bestimmte Erbkrankheit charakterisieren und von anderen trennt, ist ein wichtiger Teilschritt. Charakteristische Eigenschaften (Merkmale) von Publikationen über Erbkrankheiten können beispielsweise sein: Vorkommen von Krankheitsnamen, Name des Gens, das für die Krankheit verantwortlich ist, sowie Medikamente mit denen die Krankheit behandelt werden kann (siehe Abbildung 2).
Zunächst bildet jedes einzelne Wort in einem Text eine Eigenschaft (Merkmal).
Damit wir uns auf die Merkmale konzentrieren können, die für Text und Klasse spezifisch sind, werden vorab Stoppwörter („*and*“, „*the*“) herausgefiltert [3].

Die Basis für die Analyse der Texte bildet das Vector Space Model (VSM). Im VSM werden Dokumente als Vektoren von Merkmalen repräsentiert. Dabei bildet jedes unterschiedliche Merkmal eines Dokumentes eine Dimension des Vektors [7]. Aus der Gesamtheit der Merkmale aller Dokumente wird der Featurevektor erzeugt. Dieser ist in seiner Länge konstant und aus ihm werden die Dokumentenvektoren erzeugt.
In einem Dokumentvektor wird jedem Merkmal eines Dokumentes ein Gewicht zugeordnet. Das Gewicht ist eine Zahl, die ausdrückt, wie repräsentativ ein Merkmal im aktuellen Text ist.

Eine klassische Maßzahl ist das *tf-idf* [8] welche das Verhältnis der Häufigkeit des Auftretens eines vorkommenden Merkmals im konkreten Text zu der Häufigkeit des Auftretens in der Gesamtmenge (alle Texte) ausdrückt.
Hauptaufgabe ist es, die Gewichte der Dokumentvektoren sinnvoll zu bestimmen. Zunächst wird nur gezählt wie oft ein Merkmal im Dokument vertreten ist. Eine bessere Gewichtung könnten wir vornehmen, wenn wir zusätzlich betrachten, wo das Merkmal im Text vorkommt. Zum Beispiel könnte der Krankheitsname, wenn er im Titel des Textes vorkommt oder in der von den Autoren mitgelieferten Stichwortliste (*Keywords*) genannt wird, höher gewichtet werden. [6,7,10]

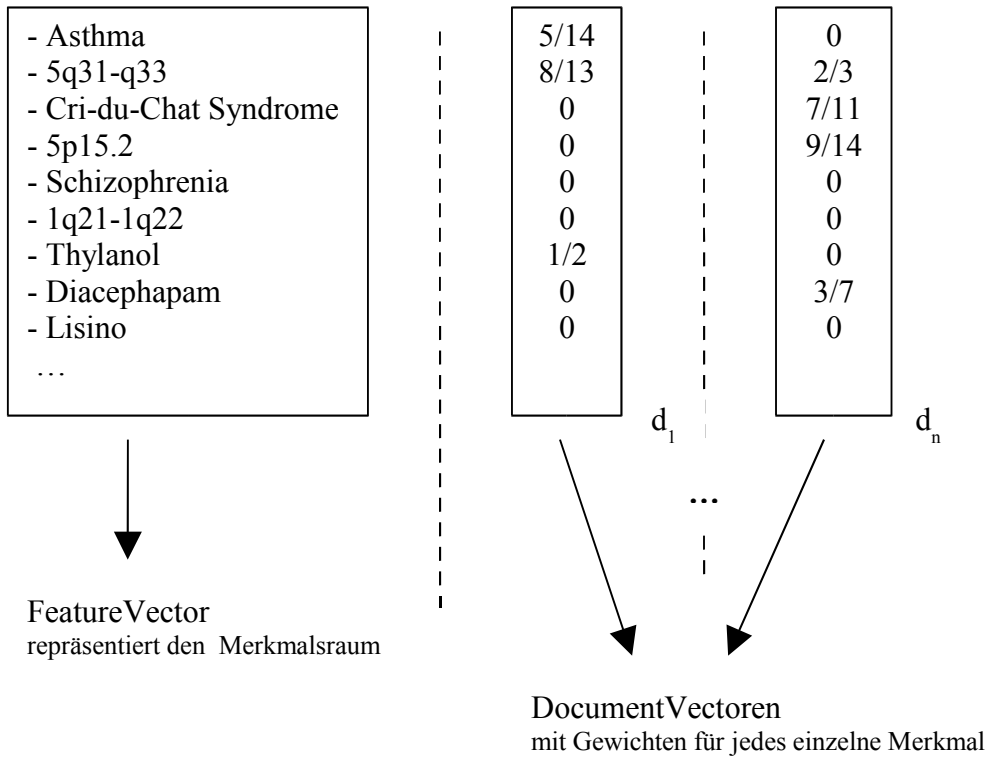


*Abbildung 2: Featurevektor und Dokumentvektor*

Für das Erzeugen der Vektoren und die Ermittlung der Merkmale stehen verschiedene Verfahren zu Verfügung. Diese sind beispielsweise Tokenization, Lemmata Tagging, Part of Speech Tagging, Shallow Parsing. Viele dieser Verfahren sind bereits implementiert worden und stehen in einer am Lehrstuhl entwickelten Klassenbibliothek zur Verfügung. [2]

Die Entwicklung des Klassifikationsmodell, welches das Kernstück des zu implementierenden Programms bildet, wird mit der Programmiersprache Java realisiert. Dabei soll durch maschinelles Lernen ein Modell entwickelt werden, dass mittels der definierten und gewichteten Dokumentvektoren erkennt, welcher Klasse der vorliegende Text zugeordnet

werden kann. Für diese Umsetzung werden zwei Verfahren verwendet: Rocchio [10] und NaiveBayes [4, 5].

Das Klassifikationsmodells wird mit den Testdokumenten getestet. Das Ergebnis ist ein Score, der angibt, welcher der 25 Klassen ein Dokument zugeordnet werden kann.
Die Korrektheit einer Zuordung lässt sich leicht bewerten, da bekannt ist, zu welcher Klasse die Testdokumente gehören (da sie anfangs von den Trainingfällen getrennt werden). Eine Bewertung des Modells wird mittels Precision / Recall – Metrik [8] sowie Cross-Validation durchgeführt und somit eine Evaluation der Klassifikation vorgenommen.

## Literatur


[1] OMIM – Datenbank
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

[2] tmlib.jar
package.wbi.textmining (am Lehrstuhl '*Wissensmanagement in der Bioinformatik*')

[3] Stoppwortlisten der Universität Leipzig
http://wortschatz.informatik.uni-leipzig.de/index.html

[4] Tom M.Mitchell. *Machine Learning.* McGraw-Hill International Editions. Computer Science Series. 1997

[5] Sang-Bum, hee-Cheol Seo and Hae-Chang Ring. *Poisson Naive Bayes for Text Classification with Feature Weighting.* Proc 6th IRAL. Sapporo, Japan. 2003

[6] Alexander Strehl, Joydeep Ghosh and Raymond Mooney. *Impact of Similarity Measures on Web-page Clustering.* 17th AAAI Conference. Austin, Texas. Pages 58-64. 2000

[7] P. Glenisson, P. Antal, J.Mathys, Y. Moreau and B. De Moor. *Evaluation of the vector space representation in text-based gene clustering.* Pacific Symposium on Biocomputing. Lihue, Hawaii. Pages 391-402. 2003

[8] Hagit Shatkay and Ronen Feldman. *Mining the Biomedical Literature in the Genomic Era: An Overview*. J Comput Biol. 10(6):821-55. 2003

[9] Berry de Bruijn and Joel Martin. *Literature mining in molecular biology*. Proc EFMI Workshop on NLP in Biomedical Applications. Nicosia, Cyprus. Pages 1-5. March 2002

[10] Shrikanth Shankar and George Karypis. *Weight adjustment schemes for a centroid based classifier.* Computer Science Technical Report. University of Minnesota, TR00-035. 2000