

Utilizing Text Mining Techniques to Analyze Medical Diagnoses

Felix Mödritscher

(Institute for Information Systems and Computer Media,
Graz University of Technology, fmoedrit@iicm.edu)

Roland Tatzl

(IT and IT Marketing, Campus 02,
University of Applied Sciences Graz, roland.tatzl@campus02.at)

Regina Geierhofer

(Research Unit HCI4MED, Institute of Medical Informatics, Statistics & Documentation,
Medical University Graz, Austria, regina.geierhofer@meduni-graz.at)

Andreas Holzinger

(Research Unit HCI4MED, Institute of Medical Informatics, Statistics & Documentation,
Medical University Graz, Austria, andreas.holzinger@meduni-graz.at)

Abstract: Due to the increasing amount of medical patient data collected in hospitals, technology-based methods are of increasing interest for processing and analyzing such materials. Therefore, computer supported techniques have to be evaluated by means of their efficiency for this application area. In this paper, we introduce an approach for analyzing expert comments on magnetic resonance images (MRI) diagnoses by applying a text mining method in order to scan them for regional correlations. Consequently, we propose a calculation of significant co-occurrences of diseases and defined regions of the human body, in order to identify possible risks for health, and we present a special tool, which we have implemented in order to test this approach.

Keywords: Information Retrieval, Text Mining, Performance, Medical Documentation

Categories: H.3.1, H.3.3, I.2.7, I.7, J.3

1 Introduction and Motivation

The application of sophisticated medical information systems amasses large amounts of medical documents, which must be reviewed, observed and analyzed by human experts [Holzinger et al., 2007]. All essential documents of the patient records contain at least a certain portion of data which has been entered in free-text fields. Although text can be *created* simple, the support of automatic analysis is extremely difficult [Gregory et al., 1995], [Holzinger et al., 2000], [Lovis et al., 2000]. In order to support the end users during their daily work, both technological performance and cognitive performance must be considered [Holzinger, 2002] and the integration of usability methods on systemic level is essential [Holzinger, 2005]. Against this background, the Institute for Medical Informatics, Statistics and Documentation (IMI) at the Medical University Graz (www.meduni-graz.at/imi) has been carrying out a variety of projects to analyze and process medical documents by applying computer-

based techniques and to present the information human centered. One of these approaches is described in this paper, which is structured as follows: Section 2 outlines some theoretical background of text mining in the field of medical informatics as well as some related projects in this area. In section 3 we present our approach and the web-based tool. Finally, we discuss the utility of our methodology on the basis of specific results, experiences, problematic aspects and opportunities. Finally, the paper is concluded by providing an outlook on further research activities.

2 Text Mining for Medical Documents

Contrary to structured information, textual information is characterized by its inherently *unstructured and fuzzy nature*, being language and domain dependent, as well as consisting of sentences and sub-sentences. Basically, text mining approaches apply statistical or pattern based algorithms in order to extract significant key-word associations or to mine for prototypical documents (e.g. for parts-of-speech tagging or term extraction). Text mining is considered a sub-specialty of Knowledge Discovery from Data (KDD) and has been headed strongly in the direction of natural language processing (NLP) during the last decade.

In accordance with [Granitzer, 06], the following stages of the text mining process can be identified: (1) Pre-processing is necessary to prepare texts, e.g. by removing layout information, or to improve the text quality by utilizing methods like stemming. (2) Information extraction comprises the stage which transforms unstructured text entities into structured elements such as database entries. (3) By applying statistical methods, features of an information space can be extracted from the structured text. Hereby, methods such as frequency analyses (e.g. of words), collocations or co-occurrences are utilized. (4) As a result, each text object is described with several features, which, for instance, spans an *n-dimensional vector* space with *n* equals to the *number of features for a text object*. This feature space can be utilized for further operations, e.g. comparing texts, visualizing relations in the text corpus, calculating similarities or rankings, etc.

In the last years, mining in medical digital libraries has come up with new findings and hypotheses. [Srinivasan, 04] reports on the development of text mining methods on the basis of medical subject headings (MeSH). These algorithms generate hypotheses by identifying potentially interesting terms related to the specific input. Additionally, new protein associations have been found by clustering learning and vector classification techniques [Fu, 03]. Another approach utilizes a rule-based parser and co-occurrence for extracting and combining relations [Leroy, 03]. Further, a new way to use thalidomide has been discovered by mapping phrases to concepts of the Unified Medical Language System [Weeber, 01]. Finally, co-occurrences are useful to build up gene networks [Jenssen, 01] and to discover gene interactions [Stephens, 01]. Derived from these experiences, three kinds of application areas for text mining can be outlined in the field of medical documentation:

Firstly, such methods are applied in order to build up an infrastructure or models for biomedicine, i.e. by finding patterns or relations in texts and generating a feature space. Inspecting the projects on the basis of the well-known text mining framework GATE (<http://gate.ac.uk/projects.html>), MultiFlora or myGRID can be identified as examples for this approach. Secondly, text mining is used to observe and retrieve

documents with innovative ideas in the scope of a restricted domain (cf. projects like BioRAT or InESBi). Thirdly, text mining techniques are utilized to extract information or features out of a medical text corpus, which is generated as product of clinical documentation for further operations such as information retrieval. The MedDictate software comprises one solution in this scope. Although this last category of applications areas overlaps partially with the first one, there are only a few reports about the mining of medical diagnoses. In addition to these experiences, we want to report on our approach towards aiming at the detection of diseases in MRT diagnoses. In the following two sections, this project and its outcome are described in detail.

3 The Solution Approach

The success of text mining methods in medical research was the origin of our idea to apply statistical techniques in order to find hints for possible locations of diseases in MRT diagnoses. Therefore, we aimed at the topological proximity between anatomic structures and pathologic expressions and implemented a tool which calculates the significant co-occurrences of anatomic and pathologic terms within the diagnoses.

3.1 Basic Algorithm and Methodology

This calculation of significant co-occurrences is based on the Poisson distribution. In accordance with [Heyer, 06], the original formula can be simplified for two different ranges of the input parameters (see also figure 1). Hereby, a stands for the number of sentences containing term A , b for the number of sentences containing term B , n for the number of all sentences and k for the number of sentences containing both terms.

Be	$\lambda = \frac{a \cdot b}{n}$	then:	$sig(A, B) = \frac{-\log\left(1 - e^{-\lambda} \sum_{i=0}^{k-1} \frac{1}{i!} \cdot \lambda^i\right)}{\log n}$
If	$\frac{(k+1)}{\lambda} > 2.5$	then:	$sig(A, B) \approx \frac{\lambda - k \cdot \log \lambda + \log k!}{\log n}$
If	$k > 10$	then:	$sig(A, B) \approx \frac{k \cdot (\log k - \log \lambda - 1)}{\log n}$

Figure 1: The three formulas to calculate significant co-occurrences of two terms

Due to performance reasons, we decided to implement the calculation of the significant co-occurrences independently, instead of re-using existing text mining modules. The different ways to calculate the co-occurrence allow the usage of the fastest algorithm for each diagnose, as the simplified formulas (the last two in the figure) require less time and processing power. However, we also had to consider pre-processing steps of the MRI diagnoses in order to complete calculations on the initial text corpus within a reasonable period of time.

3.2 Text Corpus and Pre-Processing

Accompanying measures during the evaluation of an information extraction tool revealed the need for additional assistance in finding topological relations between anatomic structures including regional indicators and diseases such as tumors. Given these requirements, there was a demand for domain specific databases for anatomic structures and pathologic expressions.

At the starting point of this project, we used a text corpus of about 6.000 diagnoses, which comprises comments of medical experts on magnetic resonance imaging (MRI) material. These findings derived from different radiologists, are completely written in capital letters and full of medical terms. Thus, we also faced the problems of synonyms, medical dialects and abbreviations. Further, the diagnoses are spread over a period of 17 years. As a consequence, time-dependent changes of terminology might be possible. These textual diagnoses were made anonymous and imported into a database. Considering the systemic performance and the sentence-based statistical calculation, the texts were also split up into sentences. Pre-processing of the free text is so realized that the occurrences of expression pairs are counted and stored. Consequently, the calculation of the co-occurrences can be executed by means of one of the three formulas which are shown in figure 1. Performance issues demanded a full-text indexing of the diagnoses as well as a reduction of the anatomic terms. Unfortunately, two or three character words are found in many other words, so they have to be excluded from the reference database. Additionally, we were in need of anatomic and pathologic terms for our approach. A corpus of approximately 6800 anatomic structures was generated from an anatomic dictionary [Dauber, 05]. This dictionary offers a rough allocation of anatomic structures to anatomic regions. More precision in finding such structures can be reached by using synonyms, the gathering of which is also time consuming. Efforts have been made to start with a synonym enhancement for the anatomic data at IMI, which has been used for the calculation. On the other hand, a Pathology database has been set up manually due to a lack of accessible resources. These corpuses represent the domain specific database sources for the statistic calculation and can be maintained in special application modules.

3.3 The Web Application

A basic user access control system is used for logging purposes. Maintenance operations must be executed as administrator. User actions include registration, editing of the registration information, login, logout and observing the login history. The application itself offers modules for the following functionality:

- Diagnoses can be listed and filtered according to two terms.
- Anatomic terms can also be filtered.
- The location for each anatomic term is indicated.
- The synonyms module shows all available synonyms for each anatomic term. These synonyms cover small parts of terms, but will increase in future.
- The menu option "ADD PATHOLOGY" provides a dialogue to add a term.
- Significant co-occurrences are listed in module COOCCURRENTS. Additionally new calculations can be started.
- A maintenance module provides splitting of the diagnoses as well as calculation of the occurrence of the single terms.

3.4 Core Functions and Advantages

For performance purposes, the administrator can initiate a pre-calculation of the occurrence of each single term. Thus the number of sentences and anatomic terms for the statistic calculation can be reduced. The splitting of the diagnoses is the second method of improving performance during the calculation. Additionally the split sentences are reduced by excluding sentences with a character length < 15. Thereby abbreviation sentences are most likely eliminated.

The screenshot shows the website of the Medical University of Graz, specifically the Institute for Medical Informatics, Statistics and Documentation. The user 'jdoe' is logged in. The search interface displays the following information:

- Diagnoses filtered by: **TUMOR** and **KLEINHIRN**
- Limit: 100
- Search for: TUMOR and KLEINHIRN
- filter
- Result count: 43

The search results are displayed in a table with columns for 'Type/Year' and 'Diagnosis':

Type/Year	Diagnosis
MR 2003 11	ZUSTAND NACH KRANIOTOMIE HOCHFRONTAL RECHTS UND TELRESEKTION EINES KELBENFLUEGELMENINGEOMS. ZUSTAND NACH RADIOCHIRURGISCHER KONVERGENZTHERAPIE DES PARASELLAEREN MENINGEOMRESTES IM DEZEMBER 2002. IM VERGLEICH ZUR VORUNTERSUCHUNG VOM 6.6.2003 BESTEHT KEINE WESENTLICHE BEFUNDAENDEUNG. UNVERAENDERTE DARSTELLUNG UND AUSDEHNUNG DER PARASELLAEREN TUMORRESTE MIT INFILTRATION DES SINUS CAVERNOSUS BIS NACH INTRAPELLAR REICHEND. NACH VENTRAL AUSDEHNUNG DES TUMORS BIS IN DIE FISSURA ORBITALIS SUPERIOR, NACH KAUDAL GEGEN DAS CAVUM TRIGEMINALE. SIE A. CAROTIS INTERNA IM INTRACAVERNOESEN VERLAUF ZIRKULAER UMSCHIEDEN UND HOCHGRADIG KOMPRIMIERT. NACH KRANIAL RECHT DER TUMOR BIS AN DAS CHIASMA OPTICUM HERAN, JEDOCH KEIN HINWEIS AUF KOMPRESSION DESSELBEN. KLEINE KORTIKALE NARBE IN DER RECHTEN KLEINHIRNHEMISPHERE. SONST ALTERSENTSPRECHENDE DARSTELLUNG BEIDER GROSS- UND KLEINHIRNHEMISPHEREN. KEIN ANHALTSPUNKT AUF LIQUORZIRKULATIONSSTOERUNG.
MR 2003 11	ZUSTAND NACH KRANIOTOMIE HOCHFRONTAL RECHTS UND TELRESEKTION EINES KELBENFLUEGELMENINGEOMS. ZUSTAND NACH RADIOCHIRURGISCHER KONVERGENZTHERAPIE DES PARASELLAEREN MENINGEOMRESTES IM DEZEMBER 2002. IM VERGLEICH ZUR VORUNTERSUCHUNG VOM 6.6.2003 BESTEHT KEINE WESENTLICHE BEFUNDAENDEUNG. UNVERAENDERTE DARSTELLUNG UND AUSDEHNUNG DER PARASELLAEREN TUMORRESTE MIT INFILTRATION DES SINUS CAVERNOSUS BIS NACH INTRAPELLAR REICHEND. NACH VENTRAL AUSDEHNUNG DES TUMORS BIS IN DIE FISSURA ORBITALIS SUPERIOR, NACH KAUDAL GEGEN DAS CAVUM TRIGEMINALE. SIE A. CAROTIS INTERNA IM INTRACAVERNOESEN VERLAUF ZIRKULAER UMSCHIEDEN UND HOCHGRADIG KOMPRIMIERT. NACH KRANIAL RECHT DER TUMOR BIS AN DAS CHIASMA OPTICUM HERAN, JEDOCH KEIN HINWEIS AUF KOMPRESSION DESSELBEN. KLEINE KORTIKALE NARBE IN DER RECHTEN KLEINHIRNHEMISPHERE. SONST ALTERSENTSPRECHENDE DARSTELLUNG BEIDER GROSS- UND KLEINHIRNHEMISPHEREN. KEIN ANHALTSPUNKT AUF LIQUORZIRKULATIONSSTOERUNG.
MR 2004 07	ANAMNESTISCH ZUSTAND NACH MENINGEOM-OPERATION LINKS INFRATENTORELL. POSTOPERATIVER PARENCHYMDEFEKT AN DER DORSALEN LATERALEN CIRCUMFERENZ DER LINKEN KLEINHIRNHEMISPHERE MIT UMGEBENDEN POSTOPERATIVEN SIGNALVERAENDERUNGEN. KEIN HINWEIS AUF REST- BZW. REZIDIVTUMOR. LOKOREGIONAER GERING VERSTAERKTES MENINGEALES ENHANCEMENT ALS AUSDRUCK VON NARBE NBILDUNG. DAS UEBRIGE HIRNPARENCHYM UNAUFFAELLIG. KEIN HINWEIS AUF LIQUORZIRKULATIONSSTOERUNG. NEBENBEFUND: TELWEISE POLYPOIDE SCHLEIMHAUTSCHWELLUNG IN BEIDEN KIEFERHOEHLN UND IN DER RECHTEN KELBENHOEHLE. RANDSTAENDIGE SCHLEIMHAUTSCHWELLUNG IN EINZELNEN ETHMOIDAL ZELLEN.

Figure 2: Filtering diagnoses according by the terms “TUMOR” and “KLEINHIRN”

Search for topological relations is based on MRI diagnoses in a heterogeneous context (e.g. cranial and spinal MRI diagnoses), as visualized in figure 2. The table of diagnoses can be filtered manually and retrieved in a list. Simple IR-techniques, such as query term highlighting, are used to visualize the results. These functions support medical experts on comparing the results for the calculated pairs of terms.

Anatomic and pathologic terms can be edited in separated dialogues. Because of a lack of a Pathology reference corpus, authoring functionality has been implemented for the application in order to manually add, modify or remove expressions. The processing of the medical free text is based on the formulas described in subsection 3.1. In order to improve the overall performance during calculation, the query considers only sentences which contain any of the anatomic or pathologic terms and only terms which were previously identified in the diagnosis corpus.

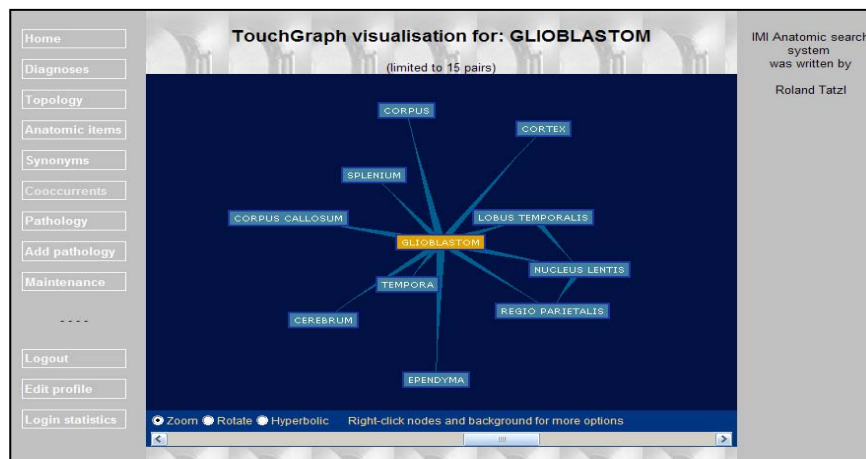


Figure 3: Tough graph visualization for the pathological term “GLIOBLASTOMA”

The resulting pair-list is shown in a table in descending order of significance for each expression. For further investigations, the most promising pairs can be used to filter the diagnoses in order to evaluate the results. For visualization purposes, a *Touchgraph* applet was implemented which enables the medical expert to estimate the proximity at a glance (see figure 3). In addition to the basic relations between the pathologic expression and the anatomic structures, the interconnection significance among the anatomic structures is calculated, to show their potential proximity within the corpus of diagnoses. This visualization clearly shows the topologic relationship.

4 Results and Experiences

The results of the calculation must be analyzed in their special context. Each pair of terms implicates specific anatomic-pathologic issues, and, in addition, the meaning of the co-occurrence must be interpreted individually.

4.1 Results for an Exemplary Scenario

The results are discussed on the basis of the malign brain cancer glioblastoma, which generally occurs in the group of middle-aged men and is located most likely in the corpus callosum and the temporal lobe. According to [Poeck, 87], glioblastoma does not occur in the cerebellum, a fact that is not contradicted in the result set at least. An analysis in a linguistic database (<http://wortschatz.uni-leipzig.de>) showed a very low occurrence in common language sources, such as newspaper articles and books, and emphasizes the importance of a well maintained domain specific database. The result set for the calculation (see figure 5) showed 10 pairs of terms for glioblastoma. We emphasize, that a highly significant co-occurrence does not prove the affliction of an anatomic structure with the paired disease. Also, a co-occurrence suggests a relation for further investigation. The most significant result suggests the co-occurrence with TEMPORA and SPLENIUM, whereby both locations are well known and located in

the most significant decile of the result set. The second group of results does not make sense, because the associated anatomic terms CORPUS or EPENDYMA (thin epithelial membrane lining the ventricular system of the brain and the spinal cord canal) are too common and do not implicate worthwhile location information. The third group showed a combination of terms which are less well known and might be an interesting hint for further investigation.

4.2 Discussion of Experiences, Problematic Aspects and Opportunities

The resulting set of the calculations are listed in descending order according to the significance. The overall statistical evaluation says that the upper decile contains the most significant co-occurrences with a high probability. Some of the less significant results include interesting combinations of terms which may point out valuable new conclusions. Finally, syntax highlighting enables the observer to find the terms of the query quite easily in the result set of diagnoses. Despite such experiences, we also identified the following problematic aspects for applying statistical text mining techniques on diagnoses: (1) The group of results which are not useful is an evidence of the weakness of the anatomic reference corpus. Expressions with a too widespread a meaning should not be considered in order to reduce the wrong results. (2) The amount of diagnoses must be increased. Examples from professional common language research show that about 5 million sentences or more are required to validate our approach. (3) The diagnoses are specific for responsible radiologists. Therefore, they are not thoroughly comparable, as different experts tend to use other terminologies. (4) Words with only few characters (like “OS” or “COR”) are not suitable for searching purposes. Thus, they have to be eliminated. (5) Acronyms and abbreviations (dotted) cause difficulties when splitting the diagnoses into sentences.

Nevertheless, we find that our methodology for analyzing medical diagnoses comprises a promising approach: The analyzing of medical text corpora is fully computer driven and fully automated. The overall calculations require from a few minutes up to some hours, depending of the computational hardware and the amount of data. Thus, this method can be applied on text corpora at any time, e.g. to use other, more accurate anatomic and pathologic expressions for old diagnoses. Secondly, our tool is of interest for clinical professionals in order to support them at their daily tasks, for instance during pre-analyzing diagnoses. We also implemented some functions to support general medical experts in their daily work in order to reduce their cognitive load (see subsection 3.3). Finally, this kind of text mining algorithm could be also valuable for other application areas.

5 Conclusion and Future Work

We emphasize the importance of computer-based methods in medical documentation and the automatization of clinical processes, including analyzing diagnoses. In this context, we developed a methodology for text mining in medical text corpora and implemented a tool to evaluate our idea. The outcome of the calculations showed valuable results although based on a relatively low number of sentences. Observing all the diagnoses, generated in a hospital daily, will definitely improve the diagnostic value. However, we still have no proof that the anatomic structure is affected by the

related disease, however, our experiences encouraged us to carry out further research efforts on these co-occurrences. Mining in large amounts of textual medical information can reveal new patterns for various questions. There will be a continued need for new mining assistants solving problems which are not even known today. We identified a huge benefit for the administration in identifying trends and developments in time to come to the appropriate decisions. The appropriate information presentation to the end users is a central future challenge, in order to keep their cognitive load in an optimum level, providing cognitive performance support.

References

- [Dauber, 05] Dauber, W., Feneis' Bild-Lexikon der Anatomie, 9th ed, Stuttgart: Thieme, 2005.
- [Fu, 03] Fu, T., Mostafa, J., Seki, K., Protein association discovery in biomedical literature, Proc. ACM/IEEE-CS Joint Conference on Digital Libraries (2003), 113-115.
- [Granitzer, 06] Granitzer, M., KnowMiner: Konzeption und Entwicklung eines generischen Wissenserschliessungsframeworks, Dissertation, TU Graz, 2006.
- [Gregory et al., 1995] Gregory, J., Mattison, J. E. and Linde, C.: "Naming Notes - Transitions from Free-Text to Structured Entry", Methods of Information in Medicine, 34, (1995), 57-67.
- [Holzinger et al., 2007] Holzinger, A., Geierhofer, R. and Errath, M.: Semantische Informationsextraktion in med. Informationssystemen, Informatik Spektrum, 30, (2007), 69-78.
- [Holzinger et al., 2000] Holzinger, A., Kainz, A., Gell, G., Brunold, M. and Maurer, H.: Interactive Computer Assisted Formulation of Retrieval Requests for a Medical Information System using an Intelligent Tutoring System, ED-MEDIA 2000, Montreal, (2000), 431-436.
- [Holzinger, 2002] Holzinger, A.: Multimedia Basics, Volume 2: Learning. Cognitive Fundamentals of multimedial Information Systems, Laxmi, New Delhi, (2002), available also in German: www.basiswissen-multimedia.at
- [Holzinger, 2005] Holzinger, A.: Usability Engineering for Software Developers, Communications of the ACM, 48, (2005), 71-74.
- [Jenssen, 01] Jenssen, T.K., Laegrid, A., Komorowski, J., Hovig, E., A literature network of human genes for highthroughput analysis of gene expression, Genetics, 28, 1, (2001), 21-28.
- [Leroy, 03] Leroy, G., Chen, H., Martinez, J.D., Eggers, S., Flasey, R.R., Kislin, K.L., Huang, Z., Li, J., Xu, J., McDonald, D.M., Ng, G., Genescene: Biomedical text and data mining, Proc. ACM/IEEE-CS joint conference on Digital Libraries, (2003), 116-118.
- [Lovis et al., 2000] Lovis, C., Baud, R. H. and Planche, P.: Power of expression in the patient record: structured data or narrative text? Int. Journal of Med. Informatics, 58, (2000), 101-110.
- [Poeck, 87] Poeck, K., Neurologie, 7th edition, Berlin: Springer, 1987.
- [Srinivasan, 04] Srinivasan, P., Text Mining: Generating hypotheses from MEDLINE, American Society for Information Science and Technology, 55, 5, (2004), 396-413.
- [Stephens, 01] Stephens, M., Palakal, M., Mukhopadhyay, S., Raje, R., Mostafa, J., Detecting gene relations from Medline abstracts, Pacific Symp. on Biocomputing, (2001), 483-496.
- [Weeber, 01] Weeber, M., Klein, H., Berg, L., Vos, R. Using concepts in literature-based discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium discoveries, Journal of the American Society for Information Science, 52, 7, (2001), 548-557.