Abstract: Although we live in the age of multimedia, most information is still only available in text format. This is especially true in medical informatics. Consequently, research in text mining is an essential area of computer science. With the aid of statistic and linguistic procedures, text mining software attempts to dig out (mine) information from plain text. The aim is to transform data into information. However, for the efficient support of end users, facets of computer science alone are insufficient; the next step consists of making the information both usable and useful. Consequently, aspects of cognitive psychology must be taken into account in order to transform information into knowledge and finally wisdom. In this paper we describe some experiences we made with text mining in medicine and present a view of current and future challenges in Research & Development in order to match aspects of both Informatics and Psychology. We are aiming to provide performance support in clinical decision making. Our end users include professionals in the area of medicine and health care. Our goal is improved presentation of information – in order to enable enhanced cognitive information processing, resulting in easier and more rapid building of knowledge.

Keywords: Semantic Usability, Information Overload, Performance Support, Information Presentation, Medical Documentation
Categories: H.3.1, H.5.2, I.2.7, J.3

1 Introduction and Motivation

Information retrieval is an important part of the daily work of medical professionals and the body of knowledge in medicine is growing enormously. Clinicians must rely on various sources of medical information and increasingly they are faced with the problem of too much rather than too little information [Slawson et al., 1994]. The problem of information overload is rapidly approaching, however, it was both interesting and motivating for our work, that most of the past research about
information overload in health care dealt with information from medical literature search [Sullivan et al., 1999], [Hall and Walton, 2004], and only a few dealt with information overload from clinical information systems [Noone et al., 1998]. Interestingly, people often compare the search in clinical databases to a classical Web search. However, the widespread format of medical documentation is strongly oriented towards individual patients and therefore the number of documents found is measurably smaller than one is accustomed to receiving from, for example, a typical Web search, although they may also comprise several hundred single documents [Holzinger et al., 2007]. Contrary to a Web information search, there is a much higher recall and precision expectancy [Baeza-Yates and Ribeiro-Neto, 2006], [Witten and Frank, 2005] from a search within the medical environment. Furthermore, it is of crucial importance that all documents necessary for the patient’s treatment are at the disposal of the medical professionals within the shortest possible time, because often – for example in an emergency – the time for viewing the data is severely limited. A simple data base and/or text search is frequently not a sufficient information system with which to support the doctors and physicians effectively.

It is a crucial factor that a large part of the information necessary to make an appropriate sample is only available in text form – the usual electronic equivalent of medical texts. Support by expert systems is therefore practically impossible without first processing those textual documents. Although, current patient documents also contain codes, among other things for diagnosis or treatments, these only represent a very small part of the actual information available. This has two main causes: the routine coding is often for primarily administrative reasons, for example, accounting or controlling; and mapping information to a standardized system of concepts usually involves a loss of information, which is unacceptable in a field where exactitude is vital. An important step, in order to analyze the texts of, for example, Medical Finding Reports to the fullest extent, is to extract the medical statement, diagnoses and other texts. These facts can be reproduced in a suitable system of concepts and be combined with the (medical) codes originating from administrative and accounting documentation. When this semantic information extraction is completed, an appropriate form of Information Presentation for the interaction between medical professionals and the system must be developed.

2 Theoretical Background: Text Analysis

The most important means of communication in patient treatment – after the spoken word – is the written word, whereby given that dictation is usual within the medical environment, the intersection between the two is, from the viewpoint of the doctors and physicians, somewhat indistinct. All essential documents of the patient records contain at least a certain portion of data which has been entered in free-text fields. Although free text can be created simply and intuitively, it makes an automatic analysis enormously difficult [Gregory et al., 1995], [Holzinger et al., 2000], [Lovis et al., 2000]. One possible solution for this is the coding and/or the use of keywords for the information contained in the free texts.

In contrast to everyday language, medical language affords a multitude of existing and controlled vocabularies, classifications and nomenclatures.
Here, medicine has a long tradition, partly with schemata which is recognized and accepted world-wide, although current ways of producing text for documents in medicine are not optimal; natural language aims to generate text from a conceptual representation of not only certain facts, but also knowledge [Huske-Kraus, 2003].

Therefore, an analysis of medical texts can not be limited to a simple classical information retrieval by means of the search for words and word components, even when they have syntactically and orthographically correct contents. It is imperative that we take into account the medical connotation of the respective term and, above all, the context, which is the information that is being presented at the moment the information need is occurring [Ruan et al., 2000]. A semantic search can be supported optimally only when these measures have been taken.

### 3 Typical Procedures

The procedures used at present comprise many different methods from medical informatics, including: 1) the use of algorithms from the field of Natural Speech Processing or Natural Language Processing (NLP), 2) Information Retrieval and 3) Machine Learning; in addition, relational data bases and Topic Maps can be used [Bohm et al., 2002].

In a first step, the structured sections of the data, available in the form of semi-structured patient documents, are processed. Here, it is possible to work with the classical methods of Data Mining. An assessment of the data quality and data plausibility is meaningful, because it is not uncommon to find that the data, even if already structured, is incomplete and/or incomprehensible. Our experience shows that at least two factors are of crucial importance: The non-intuitive organization of the user interface and the priority of saving human life over imputing data. For doctors and physicians, the priority naturally lies in medical achievements. When the quality of the data has been examined, checked for feasibility and presented in the required format, then measures are taken to process the non-structured data. In order to extract the information from this data, it is necessary to carry out the usual Text Mining steps. Texts, which come from the daily clinical routine and were not subject to a systematic, linguistic review process, require more procedures than, for example, articles from scientific magazines. Which steps are necessary, and how these must be carried out in detail, depends on the kind and quality of the raw material and the further processing and/or their intended use. It is important to notice that the majority of patient documents were never examined for linguistic weaknesses. Orthographic errors are fairly common during the clinic routine and mistakes, which can obscure the meaning to a certain extent, must be taken into consideration. Physicians in general, combine different languages within a document, in particular Latin, German and English. Purely formally therefore, a multilingual approach is necessary. Abbreviations are used very frequently: Since the medical language is a technical language, it has also its own vocabulary. However, the problem is that this vocabulary varies regionally very strongly. This strongly differing vocabulary is, in addition, regularly abbreviated, particularly in the case of internal documentation.
As long as we are dealing with relatively small sets of data, for example: many medical studies fall into this category – the automation can be considered unnecessary and the Keywords can be set manually. However, larger data sets make the utilization of tools inevitable and in a hospital, where very large data sets are the general rule, they are necessarily almost essential. Here, software solutions from the area of Text Mining can provide valuable functionality.

4 Usefulness of Medical Information

In [Slawson et al., 1994] the usefulness of medical information is expressed as

$$\text{Usefulness} = \frac{(\text{Relevance} \times \text{Validity})}{\text{Work}}$$

Wherein Relevance = 0 when the document is irrelevant to the current question, and the Validity = 0 when the document itself is invalid. Work is the work required to retrieve and interpret the information. Although Slawson’s work (and many others in this area) refer to medical literature, in our experience this also holds for medical documentation. The information retrieval unit of work is known in Human–Computer Interaction as time to perform task, which is clearly measurable [Stary and Peschl, 1998]. This includes the four stages of user interaction, query definition, database retrieval, and presentation. Appropriate user interfaces can influence this process and thereby reduce the work and increase the usefulness. Consequently, it is important during the design of such interfaces to learn about the end users, who will be using it; different types of end users require different types of interfaces and the time to perform task is dependent on various metrics, including level of skill, goals, computer literacy, frequency of use and familiarity with the domain (expert/novices) [Johnson et al., 2005], [Holzinger, 2005]. The primary goal of performance support is to provide adaptive user interfaces, which assist end users to accomplish their work accurately. They must be specifically adapted to the end users individual perception of work and methods. By addressing the end users’ needs, the quality of their work, as well as their individual performance, can be improved. The adaptation of behavior of interactive applications to individual work processes can enable the achievement of this and other goals. According to [Stary and Stoiber, 2003], user interfaces can be considered as process portals reflecting the individual views on the work in which the end users are involved. Thereby, it is essential to focus on individual end user performance, rather than on purely functional specification [Stary and Stoiber, 2003].

5 Usability of Medical Information

In order to make the information received accessible, one last step is crucial: the usable presentation of the information in order that the end users are able to transform the perceived information into knowledge. Here, we must emphasize that there are some deviations in the definitions of knowledge and here, we follow the psychological definition, which deviates from the general definition in the area of knowledge management [Schneider, 2002]. Knowledge in that respect is the result of
the interpretation of information [Holzinger, 2002]. Not until the information has been psychologically perceived, assimilated and cognitively processed and accepted in the memory – through integration into existing knowledge – can the information become knowledge.

While end users are confronted with a flood of data, the designers of the user interface must learn to differentiate between information required, which must consequently be integrated into the chain of procedure, and the (currently) irrelevant data, in order to support the cognitive development of knowledge. It is easy to acknowledge that an information presentation, which has been adapted to the medical profession, can contribute considerably towards performance support for the end users. This research into the inclusion of human factors in the development of such systems is of vital importance [Kohn et al., 1999].

One serious issue concerned with textual information is the problem of Information Overload [Eppler and Mengis, 2004], when end users are confronted with too much irrelevant information, this is similar to Memory Overload, [Koppel et al., 2005], [Nielsen, 2005]. The flood of information is actually rising constantly; more and more information becomes available within our medical information systems, while the cognitive capabilities of humans during the same period of time fail to increase to the same extent [Lewis et al., 2006]. Information overload reduces the quality of work and is a major cause of difficulties, which medical professionals experience during diagnosing.

In cognition research, one of the principles used to explain this phenomenon is referred to as Cognitive Load [Sweller, 1988]. This principle, originally based on the limitations of the human short-term memory [Miller, 1956], is probably the most prominent problem in human information processing.

Therefore, the minimization of the cognitive load, which we believe to be an important factor in the maximization of the end users cognitive performance, and facilitating the assessment of the most substantial information are the most prominent goals and remain the principle of good and accepted information systems. The form of the information presentation depends considerably on the working habits, the working environment and on the existing IT System arrangements. In order to make a holistic perspective possible, the end users workflow must be carefully examined in detail and, above all, the experiences and experimental results from real life must flow into the informatics development at systemic level.

6 Tools for the analysis of medical texts

A distinctive feature for tools designed for the semantic analysis of medical texts is their original range of primary application. It is interesting that some these tools were first used mainly in accounting and for administrative documentation. Others originate from the area of document research. Typical representatives of the first group would be "Semfinder" or "ID DIACOS", a typical representative of the second group "Morphosaurus" [Marko et al., 2005]. The knowledge base of tools, which originates from the routine, is based on years of experience with medical documentation conforming to clinical convention. Local language characteristics need to be allowed for, just as the regional legislation and naturally the rules of the respective financing system must be taken into consideration. Through this close
connection with the hospital business, the tools are particularly well-engineered and therefore particularly suitable for these areas.

The second group of tools was originally used to retrieve scientific texts. Previous to publication, these texts were reread, peer-reviewed, edited etc., of course, the correctness of the contents were also implicitly controlled.

As a result, the raw material has a substantially higher quality. Since scientific publications are, per se, intended for a wider public, they exhibit a far smaller regional specificity than the texts from the clinical routine and patient treatment. Naturally, texts retrieved by the second group offer far better conditions for key-wording etc., than texts from routine medical work. A further differentiating factor is the width of the approach. Here the implementations, which are intentionally limited to an extract of the possible documents and/or contents, are divergent to those with a more comprehensive approach. For example, on the basis of sentence fragments, "Semfinder" supports a diagnostic coding – primarily interactive but also in batch when desired – and only includes other aspects as for example: treatments, medicines and patient occupation, to the extent that this is relevant to the diagnostic coding.

Other programs focus on some special document types, for example for key-wording biomedical texts or annotating pictorial material. However, some products – such as the "ID DIACOS" linked to "ID Tasmed" – exclude no document and/or no aspect of medical documentation. The goal is to make the patient document accessible in its entirety.

The different computer-linguistic procedures, which are used for the analysis of natural speech, can also be drawn on for the differentiation of documents. These are tools that can access a specific corpus (e.g. the MorphoSaurus mentioned above) and furthermore, they can work with semantic nets and/or ontologies (e.g. ID DIACOS). However, our experience showed that a rule-based approach supplied quite good results – particularly when the area is relatively small.

7 Conclusion and Future Outlook

In the course of the further development of the organization-spanning, interlaced patient document (keyword: e-Health), it is to be anticipated that the semantic development of the medical documentation will increase in significance. The synergy with the Semantic Web group will probably continue to accelerate the development. Similar to the World Wide Web, the "World Wide Health Net" would offer the possibility of a more purposeful search. In addition to this improvement, the use of information in expert systems would actually be possible to a wider extent than ever before. A possible application here would be, for example, a "Ranking" of medical documents, according to their relevance with regard to a definite question. The steps described above – to convert a semi-structured patient document into a machine processable database – are an important advance on this path. However, a further component for the successful implementation of any health net is the expert knowledge, as to the relevance of different combinations of observations and findings for the respective medical problem. To date, no generally accepted specified set of rules to specify the characteristics of relevant findings exists.

Even for wide-spread diseases, several different sets of rules as to the correct procedure often exist – in medicine; these are also referred to as clinical paths. It is
essential to acquire the existing expert knowledge for each area considered. However, expert knowledge is not always simple to verbalize, and often not available – at least consciously. Everyone has, at some time, explained their method by phases such as: "it just looks right" or "that way feels right".

In future, the conscious knowledge of the respective relevance of medical facts combined with the articulation of unconscious knowledge which is manifested in the actual observation pattern of the medical findings will become increasingly more important. From these and other questions, which arose during our work, it is obvious that there is a need for interdisciplinary cooperation between computer scientists, psychologists and educationalists.

References


