Diversity in Medical Social Media Data: Approaches, Study and Future Challenges

Kerstin Denecke
Research Center L3S
Hannover, Germany

ABSTRACT: *Medical social media data offers an additional source of information on medical issues. Web 2.0 or Medicine 2.0, respectively, open new ways in providing and accessing this information. The medical content available is highly diverse: It can deal with diseases, medical treatments, medications and the like to which in turn different aspects can be considered. Presented content can rather deal with experiences or can provide informative insights into a topic. Making this diversity visible to a user could among others help to recognise unknown facets of a topic or help to get an overview on the information content of specific data sources. The objective of this work is to introduce the problem of diversity of medical Web content and to present a variety of methods for identifying and analysing content diversity in this data. The approaches base on information extraction technologies and involve domain knowledge. They are applied to a set of medical social media data for which the content diversity is studied. Furthermore, it will be shown how diversity information related to the subjectivity can be used for ranking to improve user satisfaction.*

Keywords: Social media, Information extraction, Medical data, Web 2.0

Received: 21 September 2009, Revised 18 October 2009, Accepted 24 October 2009

1. Introduction

Social Media Data is more and more considered a valuable information source even in medicine. Weblogs, for example, have become a popular way to share experiences and information on medical conditions, to engage in discussions and to form patient communities. Due to different influence factors such as background of an author, source (e.g., blog, forum), topic etc. the content provided in medical social media data (MSDM) is diverse to a large extent. The medical blogging community for example consists of healthcare professionals writing about their daily practise and current issues related to medicine on the one hand, and of patients providing information about health related issues and experiences on living with medical conditions on the other hand.

It is of particular interest to find possibilities to automatically analyse the content diversity of MSMD among others to improve search and retrieval, identify relevant, non-obvious aspects of a topic, or to identify differences in information content for single topics. Consider the following scenario:

A woman just diagnosed with *breast cancer* wants to learn more about the disease and related aspects. She enters *breast cancer* into her favourite search engine and receives results grouped into several clusters. First, the result set is split into *Information* and *Experiences* which in turn can be grouped in subclusters separating links to texts dealing with the *disease* itself from those dealing with its *diagnostic aspects, treatment aspects* and the like. On the other hand, the *experience* cluster groups results dealing with experiences on living with breast cancer or experiences with different treatments related to this disease. This result structure offers her the opportunity to get a general impression on the different facets of the topic. This scenario shows that diversity in medical texts has different dimensions and that it can be helpful for users to get hints to the different aspects related to their query.

This paper targets at introducing the problem of content diversity in the medical domain. Diversity dimensions relevant for analysing MSMD are collected and presented (Section 2). Furthermore, methods will be introduced (Section 5) that allow to analyse some of the specified diversity dimensions (Section 7). The methods will base on previous work on knowledge representation and information extraction from medical texts. Beyond, experiments on exploiting information on diversity in ranking are presented (Section 8). The paper finishes with conclusions and remarks on future work. The main contributions of this work can be summarised as follows:

- Introduction of diversity in the field of medicine,
- Definition of relevant diversity dimensions,
- Introduction of methods to study content diversity and diversity of information content, and their
- Application to real world data.

2. Diversity in Medical Web Content

We consider diversity a co-existence of contradictory opinions or statements (some typically non-factual or referring to opposing beliefs/opinions). In more detail, various aspects can be contradictory which leads to different aspects that can be diverse including *Time, Author, Location, Resource*, *Topic*, *Aspect considered*, *Information Content*, and *Information Type*.

The dimension *Time* is relevant in the medical domain, since people may write about specific topics only at some specific time period. For example, influenza-related entries will occur more frequently in winter than in summer. Depending on the *author* group, content differs and is influenced in various ways. Further, people trust information provided by some author groups more than those offered by others. Some information might only be relevant or valid in specific regions or content differs depending on the location of an author (dimension: *Location*). For example, assuming a specific medication is only accepted in the United States. Then, more people from the U.S. are probably writing about this topic.

A large variety of Web 2.0 tools is already available (e.g., blogs, forums, twitter). Depending on the source (dimension: *Resource*), the content provided can be different ranging from general information to personal experiences (dimension: *information type*). Users might be interested in information of one source only or might want to get an overview on the opinions expressed towards this topic in the different sources.

In this paper, the focus is on three diversity dimensions related to the content: the diversity of topic, the diversity of the aspect considered and the information content. Consider the following example: Assuming that there is a weblog post written by a patient suffering from *Depressions*. In some of his posts, he is writing about his daily life, i.e. about experiencing depressions, feeling lost and sad. In other posts, we might find information on the medical treatments and medications he is confronted with. He might also give an overview on the diagnostic aspects. The different posts have the same topic (*depression*), but they are different in the kind of aspects considered (*diagnosis, treatment, medication*) and in their degree of information (referred to as information content, i.e., information vs. experience).

More specifically, we define these three dimensions as follows: Given a topic T, topic diversity concerns the correlation between T and other concepts that are frequently used together with T. In this paper, medical concepts are considered, in particular those dealing with diseases, medical treatments or medications. A topic is highly diverse, if it co-occurs with a large number of other concepts. Even if texts have the same topic, different medical aspects can be considered. While one text rather talks about the treatment of *asthma*, others may rather focus on its symptoms. We consider this diversity dimension as diversity of the aspect considered. A text is highly diverse, when its content covers different semantic groups (e.g., symptoms, drugs, procedures).

Medical Social Media Data is generated by different user groups and with different user intents. Information ranges for example from latest news on therapeutic procedures to experiences with certain medical procedures. Due to these different intents, content of MSDM differs in the degree of information provided. In the following, we consider this diversity aspect as diversity of information content.

The information content of a text describes its proportion of *affective* content and *informative* content. Using this measure, we can distinguish between primarily *informative* and primarily *affective* texts. In an affective text the author often describes actions he performed during a day, his thoughts on treatments, diseases, medications or his feelings. On the other hand, a text is considered informative, if it contains general or disease-(and/or treatment-) specific information, or news on current research results. In Section 5, we will introduce methods enabling the analysis of various content diversity dimensions in medical texts.

3. Related Work

For analysing diversity as considered in this paper, three research areas are relevant: Content and Diversity Analysis, Topic Detection and Sentiment Analysis. In this section, work related to these topics is presented focussing on work in the medical domain.

3.1 Content and Diversity Analysis

To the best of our knowledge, diversity analysis has not yet been considered for the medical domain. Existing research considered diversity in the context of Web search, or more specifically as problem of result diversification [8],[1]. Since user queries can be ambiguous regarding their intent, diversification in this context targets finding the right balance between having more relevant results of the 'correct' intent and having more diverse results in the top positions. The applied diversity measures and diversification objectives are limited to result diversification in the context of Web Search. The notions of diversity that are handled are still restricted to certain kinds of general content or category similarity, though a large range of more specific types of diversity exist. In addition to existing work, we intend to consider additional notions of diversity and introduce methods for analysing the different diversity dimensions of medical Web content.

Also related is work in the field of content analysis. Existing work focuses among others on identifying and quantifying structural and functional properties of blogs or blogger characteristics or on the usage of content. Other research measured the quality of medical Web content or studied Web search behaviour for medical content. Eysenbach examined health-related searches on the World Wide Web[7] and showed that the posted queries are very general. Therefore, it is important to provide facilities that show the diversity of query results to improve the awareness of aspects of a query.

3.2 Topic Detection

In this paper, among others topic diversity is analysed. For topic detection, different topic representation and detection methods are available, such as clustering of documents based on extracted keywords or substituting topic identification with a lexicon look-up to determine product names, person names etc. Topic models as introduced by Blei et al. consider documents as mixture of topics [3].

A shortcoming of these approaches is that also terms not related to a topic could be selected as topic term. In our work, we are only interested in medical topics. Therefore, we decided to choose another approach. Often, named entities and topic terms are considered most relevant for detecting the topic [10]. Instead of considering general named entities as topics, our approach bases on domain knowledge from the medical domain. Medical concepts are extracted from the text under consideration and the most relevant concepts are chosen as topics. Therefore, only topics related to medicine are identified, which is crucial when analysing medical content diversity.

3.3 Sentiment Analysis

Distinguishing *informative* from *affective* postings as considered in this paper is similar to the problem of subjectivity analysis. Wiebe and Riloff [13] perform such analysis by means of a Naive Bayes classifier and lexical and contextual features (e.g., subjective and objective patterns or subjectivity clues). Our approach differs by exploiting the proportion of *affective* and *informative* content for classification purposes, and specifically targeting medical blogs.

Ni et al. presented in [12] a machine-learning algorithm for classifying *informative* and *affective* articles among weblogs. Their approach differs from ours in the features exploited: They use words as features while our approach exploits medical concepts, and opinionated words.

In this work, a lexical resource for opinion mining, i.e. the SentiWordNet, is exploited to identify and quantify the information content. SentiWordNet [6] provides for each synset of WordNet[1] a triple of polarity scores (positivity, negativity and objectivity) whose values sum up to 1. It has been created automatically by means of a combination of linguistic and statistic classifiers and consists of around 207000 word-sense pairs or 117660 synsets. Existing work exploits this resource mainly for identification of opinionated words. Devitt and Ahmad [5] identify by means of SentiWordNet sentiment-bearing words in a document whose frequencies are in turn used for classification.

4. Material

For the analysis in this paper, various sources of social media data related to medicine and health have been collected.

Weblogs A medical weblog might deal with diseases, medical treatments, medications or health care politics. For our experiments and evaluations, a set of different medical weblogs written in English and all their posts have been crawled. The 4343 patient-written weblog posts have been selected randomly by collecting addresses of weblogs from the two (medical) weblog search engines Medworm and Medlogs. 1137 Physician-written blog postings have been collected from WebMD[2] only.

[1] http://wordnet.princeton.edu/
[2] http://www.webmd.com

Drug reviews for 630 drugs have been collected from Drugratingz.com (in total 3731 reviews). Users can anonymously rate drugs in several categories, including effectiveness, side effects, convenience and value; they can post and read comments. These comments, dealing with symptoms and side effects, provide the data set for our analysis while the ratings remain unconsidered.

Q&A portal Additionally, 913 query and answer postings were collected from the Mayo Clinic question answering page[3]. Everyone can post queries to this portal, but only registered physicians are allowed to reply.

Encyclopedia For our data collection, 2777 articles from Yahoo! Encyclopedia have been collected. They deal with different topics related to illnesses, treatments and drugs.

5. Method

The diversity dimensions we will study in this paper are the *content diversity* and *diversity in information content*. The proposed methods will be described in the following sections in more detail. They rely upon a document representation by medical concepts which is produced by the mapping algorithm MMTx [2]. This algorithm is based on natural language processing techniques and maps natural language to concepts of the UMLS Metathesaurus (UMLS[4]). The UMLS is a biomedical terminology that consists of around 1.7 Million biomedical concepts and integrates several biomedical vocabularies such as SNOMED CT or MeSH. Each concept defined in the UMLS is assigned to at least one of the 135 specified semantic types. The semantic types are grouped in turn into 15 main groups.

In our approach, MMTx provides for a document a list of UMLS concepts which is exploited in our diversity analysis methods. These methods in turn base upon several assumptions (see Table 1) that resulted from our previous content analysis studies.

ASS0: Information describing *diagnoses / symptoms, procedures* and *medications* is relevant medical content.
ASS1: A main topic a text deals with is related to a set of subtopics.
ASS2: The main topic of a medical text deals with a disorder, a treatment or a medication.
ASS3: A text collected from MSMD consists of medical aspects (i.e., facts), opinionated parts (i.e., experiences) and other parts (e.g., verbs, non-opinionated terms).

Table 1. Assumptions for the Diversity Analysis Methods

5.1 Analysing the Content Diversity

The main idea behind our approach is that the diversity in content is reflected by the extracted medical concepts, their semantic types and main groups. For this reason, our method mainly studies the variety in these items.

The concept-level diversity (referred to as $div_{concept}$) provides information on the variety of concepts used in a text. Therefore, we determine $div_{concept}$ by comparing the number of different concepts used in a text (co_d) to the total number of extracted concepts co (see Formula (1)).

$$div_{concept} = \frac{co_d}{co} \tag{1}$$

A $div_{concept}$ value close to zero indicates that the same concepts are used several times, i.e. the diversity of concepts is small. Given a large concept diversity in a text indicates that many different aspects are covered, while a small concept diversity shows a restricted coverage of different aspects.

[3] http://www.mayoclinic.com
[4] http://www.nlm.nih.gov/research/umls

Furthermore, we calculate the concept diversity for single main groups *Disorders, Procedures* and *Chemicals and Drugs* by considering in formula (1) only the frequency of concepts of one of these three main groups.

On a more abstract level, content diversity can be seen as variety in semantic types and main groups. We refer to this as diversity of aspect considered. Therefore, we use the proportion of the different semantic types covered by a text as measure of this kind of diversity. In more detail, formulae (2) and (3) are applied: The formulae determine the proportion of different semantic types (main groups) contained in a text on the overall number of possible types (or groups). The UMLS provides 135 different semantic types and 15 main groups. A value close to 1 indicates a high diversity, while a value close to 0 corresponds to a small diversity.

$$div_{type} = \frac{types}{135} \tag{2}$$

$$div_{group} = \frac{groups}{15} \tag{3}$$

A large concept diversity does not necessarily imply a large diversity in aspect considered and vice versa. For example, a post could contain only a few different concepts, but these concepts could belong to different semantic types and main groups, i.e., the post deals with several aspects. In this case, the concept diversity is small, but a high diversity in semantic types and main groups would be detected.

More deeper insights into the diversity of content or topic is provided when studying correlations between the topic concept and other concepts across documents. Here the assumption (ASS1) is used that a text deals with a main topic to which in turn a set of subtopics is related. Therefore, topic-diversity is studied as the variety of concepts that cooccur with the topic concept of a document. For this purpose, the representation of texts by UMLS concepts (see above) is used to (1) determine topics of texts and (2) to identify co-occurring concepts to study topical diversity.

Based on the assumption (ASS2) we determine the most frequently used concept dealing with disease, procedure or drug as topic. Given a document collection, we receive a list of topics together with the documents for which this topic has been determined. Documents with a joint topic concept are considered in the next step when concept cooccurrence pairs are determined for each topic concept. A pair is considered relevant when it occurs at least twice in one document and in at least ten documents with this topic. This restriction has been made to identify only relevant pairs of concepts. The final result is a set of concept pairs for each topic concept.

The concept pairs provide information on how diverse a topic is: If for one topic a large amount of pairs is identified within one document collection, this topic is highly diverse in this collection. In case a topic-describing concept co-occurs only with a few other concepts frequently, its diversity is rather low, i.e. only a few additional aspects are of interest to this topic. By means of the measures presented in this section, content diversity can be assessed automatically from various views: on concept-level, aspect-level, or topic-level.

5.2 Determining the Information Content

In order to study the diversity of information content, we introduce an approach that can be used to first, distinguish *informative* and *affective* texts, and second, to quantify the information content based on the assumption ASS3. We consider the medical content of a text as *informative* content which can be quantified by the number of (relevant) medical concepts. Only concepts describing *disorders, procedures* and *medications* are considered relevant (ASS1). To extract opinionated parts of a text, words that are neither medical content nor stop words are looked up in SentiWordNet (see Section 3.3). Words, for which a SentiWordNet entry exist with an objectivity smaller than 1 (i.e., at least the positivity or negativity are larger than 0.5) are considered opinionated. In addition, each word is looked up in the General Inquirer[5] to have an additional source of opinionated words. In this way, the number of opinionated words is calculated.

Finally, we can determine the degree of informative content ($degree_{inf}$) and the degree of affective content ($degree_{aff}$):

[5] http://www.wjh.harvard.edu/~inquirer/

$$degree_{inf} = \frac{co_d}{words} \quad (4)$$

$$degree_{aff} = \frac{op_d}{words} \quad (5)$$

with cod as the number of extracted medical concepts of the document d and opd as the number of opinionated words in d, and words as the number of words in d.To decide whether a text is rather *informative* or *affective*, we use a set of features consisting of number of words, number of stop words, degreeinf, degreeaff. This feature set is exploited by a supervised machine learning algorithm. Through experiments with different possible algorithms implemented in the WEKA library [14], the SimpleLogistic classifier has been chosen for classification since it performed best.

6. Evaluation

In this section, the focus is on the quality of the topic detection (see Section 6.1) and information content analysis (see Section 6.2). Mapping from natural language text to UMLS concepts using MMTx has already been evaluated among others in [4]. In section 8 we evaluate how the consideration of information content in ranking influences user satisfaction in retrieval.

6.1 Evaluation of Topic Detection

The objective of the evaluation of the topic detection approach is to analyse its quality in selecting the right concept as topic. For this purpose, the topics that have been automatically assigned to randomly selected texts of the introduced data material, were manually evaluated by a human with medical background. In more detail, assigned topics for 200 WebMD weblog posts, 100 texts from the Q&A dataset and 100 texts from patient-written weblogs are assessed. For the detected topic per text the evaluating person decides whether the concept reflects the document's topic. If not, the evaluator could assign a topic manually or mark posts whose topic is rather general, i.e. those that are not dealing with diseases, procedures or medications.

An accuracy between 87.6% and 92.7% could be determined. In the WebMD dataset lots of posts deal with nutrition, or clinical trials or health care politics and were marked general. For the other two data sets the number of posts that have been marked as general is significantly smaller.

The lowest accuracy of 87% was achieved for the patient-written postings. A reason for this is that the authors of patient-written postings sometimes don't use the concrete medical terms, but rather describe the diseases or symptoms. This leads to an increased complexity for detecting topics since mapping to UMLS concepts will fail. Nevertheless, these results show that even this simple approach is successful in assigning topics describing the content of a medical text.

The topic detection fails when there is no single main topic in a text, i.e. different topics are mentioned. In these cases, the single concepts occur with a similar frequency and a most frequent concept can not be detected. The algorithm selects then a concept randomly which leads in some cases to errors. Additional sources of error are related to the mapping to UMLS concepts which might be wrong. Further, for very short texts such as the drug ratings the algorithm fails since concepts often occur only once.

6.2 Evaluation of Information Content Analysis

In section 5.2, the approach for classifying text into one of the categories *informative* or *affective* was presented. We evaluate this approach by manually classifying 459 weblog posts from physician-written blogs. The evaluation material consists of 188 *affective* (41%) and 271 *informative* (59%) texts. The objective of the analysis is to determine the classification quality. The evaluation was performed as 10-foldcross validation. The SimpleLogistic Classifier [14] performed with 86.5% accuracy.

Table 2 shows the evaluation results for the two different classes. *Informative* texts can be slightly better identified than *affective* texts. The recall for classifying *affective* texts is significantly lower than for *informative* texts.

Category	Precision	Recall	F-Score
affective	0.902	0.786	0.84
informative	0.842	0.93	0.884

Table 2. Evaluation Results Information content classification.

The classification method fails when a text contains as much *affective* content as *informative* content. To overcome this limitation, we redefined the classification problem and determine the extent to which a text contains *informative* and *affective* aspects. In future work, the quality of this approach needs to be assessed.

7. Analysis of Diversity

The introduced measures and methods are applied to our dataset to assess and study its diversity. The results are reported in the following.

7.1 Diversity of Topics

For studying the diversity of topics in our data collections, we determine the number of different topics that is identified for each collection as well as the average number of co-occurring concepts per topic and collection. Taking into account the different data set sizes, the largest number of different topics is given by the collection of physician-written posts (385 topics) and the encyclopedia articles (787 topics). Nevertheless, the topics in the encyclopedia articles are more diverse: In average, 60 co-occurring concepts are determined for each topic. For topics in patient-written posts, 43 co-occurring concepts were identified and for the other data sources the values were even smaller with up to 14 co-occurring concepts.

Even more interestingly is the analysis of topics across datasets. From all five data sets, 1832 different topics were identified from the complete data collection. Only 18 topics are shared by all five datasets. The majority of them is related to diseases: *Adverse Effects, Sexual intercourse, Pain, Headache, Hypersensitivity, Patient currently pregnant, Depressed mood, Hemorrhage, Stress, Diabetes Mellitus, Non-insulin-Dependent, Edema, Coughing*. Three topics are related to medications: *Antibiotics, Pill, Hormones.*. Another three datasets share topics dealing with medical procedures: *Cholestorol measurement test, Chemotherapy-Oncologic Procedure, Exercise Pain Management*.

For the 18 shared topics, we analyse the overlapping aspects, i.e., which sources share most often co-occurring concepts. When considering co-occurring concepts that are common in three data sources, it is interesting to see that texts from Encyclopedia articles, Q&A portal and patient weblogs share the largest amount of aspects (57 concepts are common for the 18 topics). The drug reviews and

physician-written posts share aspects to a very limited extent (only 26 joint aspects for the 18 topics). In contrast, the data source pairs [Encylopedia and Q&A], [Drug reviews and Patient blogs] and [Physician blogs and Patient blogs] share around 100 common aspects each. The largest overlap was determined for Encyclopedia articles and Patient blogs: 300 aspects were in common.

For the topic *Diabetes Mellitus, Non-Insulin-Dependent* we studied the diversity in more detail. From the patient-written blogs, 788 co-occurring concepts could be detected that belong to 97 different semantic types. In this dataset, the topic *Diabetes* is highly diverse -a lot of different aspects are considered. In contrast, for the other data sources significantly smaller numbers have been determined (e.g. only 60 related concepts of 25 different types were identified in the physician-written posts). We conclude, that with respect to this topic, the patient-written dataset covers a broader range of related aspects than the other data sets.

These results show that the resources of our data set consider very different aspects to a certain extent. In particular, texts of the Q&A data set consider other aspects compared to the other resources. Interestingly, the encyclopedia and the patient-written posts take similar aspects into account regarding the 18 topics that share the data sets.

7.2 Diversity of Aspect Considered

In this section, the diversity of aspect considered is studied (see Table 3). The divconcept for the encyclopedia data set ('Ency' in the table) is significantly smaller than the ones for the other datasets. Obviously, the same concepts are used several times in this dataset or in the other datasets more different concepts are used, respectively. The largest concept diversity is determined for drug reviews ('drugs' in the table). These texts are rather short, and concepts are therefore less often repeated.

In contrast, the diversity of semantic types and main groups is for the encyclopedia data set much higher than for the other data sets, meaning that a broader range of aspects is considered. In the weblog data sets ('Patient', 'Phys') and the Q&A dataset ('Q&A'), the considered aspects are more restricted. The smallest diversity values in types and groups have been determined for the drug reviews. They mainly deal with disorders and medications which explains the reduced diversity in aspects considered, and also the large diversity in concepts. We can conclude that the concepts extracted from the encyclopedia

data set belong to more different semantic types, i.e. a larger variety of thematic aspects is covered. Nevertheless, the values for divtype are quite small for all five data sets. From the 135 possible UMLS semantic types only one fourth or one third is covered by the texts (when considering all semantic types). Mainly semantic types that are unrelated to diagnoses such as *Clinical Attribute, Experimental Model of Disease, Genetic Function, or Molecular Function* remain uncovered.

Measure	Patient	Phys	Ency	Drugs	Q&A
$div_{concept}$	0.76	0.70	0.52	0.88	0.68
div_{type}	0.23	0.27	0.37	0.1	0.28
div_{group}	0.68	0.78	0.87	0.4	0.83
$div_{concept}(DISO)$	0.44	0.45	0.53	0.49	0.45
$div_{concept}(P\ ROC)$	0.18	0.21	0.20	0.04	0.18
$div_{concept}(CHEM)$	0.30	0.33	0.26	0.4	0.37

Table 3. Diversity values when considering all semantic types

The concept diversity for the category *Disorder* is similar for all data sets. For the concept diversity in Procedure concepts, the smallest diversity could be ascertained for drug ratings. For the Encyclopedia articles the diversity in disorder-related concepts is higher than the one for blogs and Q&A texts. This shows that the spectrum of covered diseases in these articles is higher than in blog posts and Q&A postings. The highest diversity in concepts related to medications is determined for drug ratings which is understandable, since they are dealing with drugs and experiences with them only.

In summary, encyclopedia texts do not contain so many different concepts, but the concepts cover a large variety of different aspects in terms of UMLS semantic types and main groups. In contrast, drug reviews have a large variety in concepts, in particular related to disorders and medications. Since they focus on these two aspects, they are not diverse with respect to other aspects. Diversity of Q&A postings and weblogs lies somewhere between these two extremes.

7.3 Diversity of Information Content

Finally, we study to what extent the different resources are informative and affective and positive and negative opinions are expressed.

For the weblog data sets (patient-and physician-written), the degree of informative content degreeinf is smaller than for the other data sets. This means, compared to the number of words, less concepts dealing with disorders, procedures and medications have been identified than from the other data sets. This can be due to a more restricted use of medical concepts in weblogs. Another explanation is that concepts remained unrecognised because they have been described in common language.

For drug reviews, the degree of affective content degreeaff is higher than for the other data sets. Clearly, people express their opinions in drug reviews and therefore, they use opinionated words more frequently. The smallest value for $degree_{aff}$ has been determined for the encyclopedia and Q&A data sets.

The classification results for *informative* and *affective* reflect the natural assumption that blogs and drug reviews are more *affective* than encyclopedia articles and Q&A postings. More than 94% from the encyclopedia articles and from the Q&A postings were labelled *informative* and only around 60% of the blogs and drug reviews. The percentage of *positive* and *negative* words is almost balanced for all data sets.

Further, the information content of the texts with joint topic is analysed across resources. We will focus here only on five topics (*Headache, Adverse Effect, Pain, Depressed mood, Diabetes mellitus*), since the largest number of documents per source belong to these topics. Table 4 shows the average values for the information content of these five topics.

Interestingly, all the texts with topic *Diabetes* were labelled *informative*. The proportion of opinionated words is quite small compared to the value of the other diseases. More positive opinionated words are used than negative ones. Thus, we conclude that the *Diabetes* texts provide mainly information. In contrast, the topics *Headache* and *Depressed mood* are more negatively discussed with more opinionated words. Further, highly opinionated is the topic *pain*. In summary, this study shows that the various topics covered by all data sources are described in different ways, i.e. described more or less opinionated, and more or less informative.

	Diabetes	Depressed mood	Headache	Adverse Effects	Pain
no.texts	206	97	37	127	227
$degree_{inf}$	0.233	0.24	0.26	0.23	0.21
$degree_{aff}$	0.11	0.19	0.20	0.15	0.21
pos	57%	39%	38%	46%	45%
neg	43%	61%	62%	54%	55%
inf	100%	72%	67%	80%	70%
aff	0%	28%	33%	20%	30%

Table 4. Information content diversity for specific illnesses

8. Information Content for Ranking

We now want to use information content for adapt the ranking of search results. Assuming that users are more interested in facts than in affective descriptions, texts that are mainly *informative* could be ranked higher than *affective* posts. We evaluate whether such a modified ranking improves the user satisfaction.

For the evaluation, thirty queries have been selected from studies on healthrelated web searches (http://www.guideline. gov/search/stored queries.aspx, [11]). The average query length was 2.6 words per query. The search database comprised 18768 physician-written weblog posts. For each query, the top 5 posts collected by a standard Lucene search engine with a scoring function were collected. For our experiments, the scoring function was modified by a boost factor which was multiplied into the TFxIDF score of hits with documents labelled *informative*. For comparison reasons, (1) a Lucene search engine with TFxIDF based scoring function without boost factor and (2) a search engine that multiplies the boost factor when a document was classified *affective* were exploited to produce result sets for the queries.

We evaluated our ranking algorithm with twelve subjects who had to rate the resulting posts with respect to relevancy to the query as (0) irrelevant, (1) relevant, (2) highly relevant to the given query. The results of all three engines were shuffled. Each tester had to assess about 178 documents for all thirty queries, being neither aware of the algorithm, nor of the ranking of each assessed post. The ranking results of the engines are compared based on user assigned ratings. The quality of each ranking was assessed using the normalised version of discounted cumulative gain (DCG) [9]. All results were tested for statistical significance using T-tests.

The results show that for queries with more than hundred answers the calculated normalised DCG of 0.857 is significantly better for the adapted scoring function where *informative* texts are boosted than the corresponding value for the normal scoring function. For queries with less than hundred results, confidence of the improvement could not be verified. The average DCG value for the adapted scoring function where *affective* texts are boosted is with 0.781 significantly lower than the value for the *informative* boosted ranking (95% confidence). This result confirms the assumption that users are more interested in informative texts than in affective texts.

We conclude, that the identification of relevant texts with regard to a specific query out of a large set of texts matching the query can be improved, when *informative* texts are ranked higher than *affective* texts.

9. Discussion

The diversity study, described in this paper, shows that depending on author and source content of MSMD can be highly diverse. Some topics are discussed in more detail, others are described in a rather opinionated way. Until now, these aspects remain hidden to a user. The introduced methods would allow to present this diversity to the user. In this paper, visualisation of diversity remained unconsidered, but is foreseen as future work.

It is still a challenge to examine the quality of the introduced measures to study content diversity. Measures to quantify the quality of a diversity measure are still missing. A possibility might be to study the user satisfaction when using the measures in a medical Web retrieval engine as we did in this paper for the information content. A similar evaluation for the other diversity

measures remains open for future work. The quality of the introduced method to study diversity of the aspect considered and the diversity of topic highly depends on the quality of mapping natural language to UMLS concepts. If this mapping contains wrong concepts the calculated measures could become incorrect.

In this paper, we considered content diversity in medical social media data. In order to apply the introduced method to documents of other domains, the underlying domain knowledge has to be replaced. If such a domain knowledge is unavailable (which is certainly the case for the most domains), the problem becomes more difficult, since domain-specific topics and related aspects need to be discovered automatically. One possibility is to select nouns as topics and to look for co-occurrences with the topic noun. Another possibility is to apply topic clustering approaches such as LDA [3]. For both approaches, semantic information on detected topics as it is provided in the presented approach through semantic types and main groups is missing. In future work, we will work towards this direction to come up with a more general approach. Nevertheless, the approach presented here can be considered as baseline when testing other topic detection algorithms in the medical domain as basis for analysis of content diversity.

In section 8, we showed how the information content can be successfully used for ranking purposes. An open question is how to represent diversity to a user in an easy understandable way. The diversity measures could also be exploited for ranking search results, i.e., as mean for result diversification. Depending on the user interest and preferences, results with a large content diversity could be ranked in higher positions than those with a small content diversity.

In a diversity-aware search engine that relies upon the approach introduced in this paper, the document collection on which the search engine bases, needs to be mapped to medical concepts in advance since the mapping takes some time. The calculation of the diversity measures can be realised very fast. Also, the counting of positive and negative words could be done in advance to reduce processing time during the search process. Nevertheless, the concept representation of the texts could support the search process, among others by considering synonyms and language variations: The mapping to concepts leads to a normalisation of the language which facilitates the retrieval.

10. Conclusion

This paper introduced the problem of diversity in medical Web content. In particular, diversity in topic and information content has been considered. We described how results of entity extraction together with a domain ontology can be exploited for studying these aspects and applied the methods to a data collection of medical texts.Having information about content diversity offers the opportunity to get insights into relevant aspects related to a topic that could for example be presented to a user. Furthermore, it could help to improve the retrieval of documents: Even if topic terms are not used explicitly, relevant documents can be found based on related terms that can be automatically used to expand a user query. A third application includes the recommendation of documents that consider the same topic but different aspects. Physicians might be aware of related aspects, but patients' background knowledge on medical issues is often limited.

References

[1] Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S. (2009). Diversifying search results. *In*: WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining, p. 5–14, New York, USA. ACM.

[2] Aronson, A. (2001). Effective mapping of biomedical text to the umls metathesaurus: The metamap program, *In*: Proceedings of the AMIA Symposium, p. 17–21.

[3] Blei, D. M., Ng, A. Y., Jordan, M. I., Lafferty, J. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.

[4] Denecke, K., Bernauer, J (2007). Extracting specific medical data using semantic structures. *R. Bellazzi, A. Abu-Hanna, and J. Hunter (Eds.):* AIME 2007, LNAI 4594, Springer-Verlag Berlin Heidelberg, p. 257– 264.

[5] Devitt, A., Ahmad, K. (2007). Sentiment polarity identification in financial news: A cohesion-based approach. *In:* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, p. 984– 991. Association for Computational Linguistics.

[6] Esuli, A., Sebastiani, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. *In*: Proceedings of LREC 2006 -5th Conference on Language Resources and Evaluation.

[7] Eysenbach, G., Kohler, C. (2003). What is the prevalence of health-related searches on the world wide web? qualitative and quantitative analysis of search engine queries on the internet. *In*: Proceedings of the AMIA Annual Symposium, p. 225–29.

[8] Gollapudi, S., Sharma, A. (2009). An axiomatic approach for result diversification. *In*: WWW '09: Proceedings of the 18th International Conference on World Wide Web, p. 381–390, New York, NY, USA, 2009. ACM.
[9] Jaerwelin, K., Keklinen, J. (2000). IR evaluation methods for retrieving highly relevant documents, *In*: Proceedings of the 23rd International ACM SIGIR Conference on Research and Development in Information Retrieval.
[10] Kumaran, G., Allan, J (2005). Using names and topics for new event detection. *In*: *HLT '05:* Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p. 121–128, Morristown, NJ, USA. Association for Computational Linguistics.
[11] Mueller, H., Boyer, C., Gaudinat, A., Hersh, W., Geissbuhler, A. (2007). Analyzing web log files of the health on the net honmedia search engine to define typical image search tasks for image retrieval evaluation. K. Kuhn et al. (Eds.) Medinfo 2007*,* IOS Press.
[12] Ni, X., Xue, G., Yu, Y., Yang Q. (2007). Exploring in the weblog space by detecting informative and affective articles. WWW '07: Proceedings of the 16th International Conference on World Wide Web.
[13] Wiebe, J., Riloff, (E). Creating subjective and objective sentence classifiers from unannotated texts. CICLing 2005.
[14] Witten, I., Frank, E. (2005). Data mining: Practical machine learning tools and techniques. 2nd edition. Morgan Kaufmann, San Francisco.