Assessing Content Diversity in Medical Weblogs[*]

Kerstin Denecke

L3S Research Center,
Appelstr. 9a, 30167 Hannover, Germany
denecke@L3S.de

Abstract. In this paper, we are considering weblogs focusing on medicine and health. Given this general topic, posts can deal with diseases, medical treatments, medications and the like to which in turn different aspects can be considered. Within a diversity-aware medical search engine knowledge about this kind of diversity can support grouping of search results into diversity dimensions covered by a text. Therefore, a method is introduced for studying topic and content diversity. The approach bases on information extraction technologies and domain knowledge and is applied to a set of medical weblog posts. The diversity of topics described in this dataset will be studied in more detail.

1 Introduction

Weblogs, or blogs, have become a popular way to share experiences and information, to engage in discussions and to form communities. In this paper, we are considering medical weblogs, i.e., blogs whose topical focus is health and medicine. The medical blogging community consists of healthcare professionals writing about their daily practice 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. Therefore, the diversity of content varies a lot. It is of particular interest to find possibilities to automatically analyse the content diversity in these blog posts to enable better search and retrieval facilities. Consider the following scenario:

A person searching with the query *breast cancer* might be interested in the disease itself, in possible treatments, in medications and so on. Other dimensions are for example the content type (e.g., experience vs. information), the author (e.g., physician, patient) or the polarity. A user might be interested in experiences of persons living with *breast cancer*, or in experiences of physicians who treat patients suffering from *breast cancer* etc.

Current medical weblog search engines such as Medlogs or Medworm[1] only list search results matching query keywords in a flat list. Sometimes, results can be restricted to posts of specific author groups (e.g., physician, patients). But, additional content dimensions such as aspects considered or expressed sentiment

[*] This work is funded by the EU under 231126

[1] http://www.medlogs.com, http://www.medworm.com

remain hidden in the posts. Having methods in hand that allow to analyse and detect the different dimensions would help to present search results according to these dimensions. The work presented in this paper targets towards analysis of diversity in medical weblog posts. The focus is on analysing the diversity of content, in particular the topic diversity and the diversity of the aspect considered.

The remainder of the paper is structured as follows. Section 2 presents related work. Then, we give an overview on relevant diversity dimensions in medical texts (section 3). Then, methods and measures to study topic diversity in medical texts are introduced (section 4). This is applied to a real world data set in section 5. The paper finishes with conclusions and remarks on future work.

2 Related Work

Diversity of search results in text retrieval has been considered as problem of result diversification, i.e., finding the right balance between having more relevant results of the 'correct' intent and having more diverse results in the top positions. Existing approaches to this problem combine measures of diversity and similarity to improve the recommendation diversity. In order to improve user satisfaction, the top N search results are either ranked by diversity [1, 4, 9, 7] or diversified by clustering them according to the different diversity dimensions covered [12].

Clustering of search results is be done within the search engines Newssift and Fairspin[2]. Within Newssift, content from major news and business sources are grouped into high-level categories such as Business Topic, Organizations, Place, Person and Theme. It leverages semantic technology, but relies also on manual work. FairSpin collects all the latest news and opinion from across the Web and organizes them by political bias based on community votes.

In faceted classification, a set of category hierarchies is built [8, 11]. These capture the different facets, i.e., dimensions or features, relevant to the collection. Facets can for example be derived based on WordNet or Wikipedia as it is shown be [6]. The work presented in [10] describes several faceted search algorithms and employs collaborative filtering and personalization techniques to customize the search interface to each user.

In contrast to existing work, we intend to determine and analyse diversity in more detail and automatically. Instead of providing hierarchical browsing facilities, a potential application would be the grouping of search results according to diversity dimensions. Presenting different topical aspects to a user would enable him to see different aspects of his query at the same instant and to get deeper insights into all the facets of the topic under consideration. Furthermore, we are focusing on medical weblog posts since this is a very interesting domain from which many people can benefit. For the medical domain sophisticated ontologies exist. We will show how these can support analysis of topic diversity.

In this paper, we study the topic diversity of texts, considering a document as a mixture of topics. Topic models as introduced by Blei et al. also consider

[2] http://www.newssift.com, http://fairspin.org/

documents as mixture of topics [3]. Each topic is represented by a set of keywords together with a probability indicating the word's contribution to the topic. In our approach, topics are considered medical concepts that are provided by a medical ontology. While within clustering approaches such as Latent Dirichlet Allocation [3] automatic labeling of clusters is difficult, the use of domain knowledge helps to label document clusters and describe topics with concrete concept names.

3 Diversity Dimensions in Medical Weblogs

In this paper, we focus on processing medical weblogs written in English. A **medical weblog** deals with diseases, medical treatments, medications or health care politics, i.e., its main topic is medicine or health care. They can be differentiated with regard to their author into blogs written by health care professionals and written by non-healthcare professionals [5].

Diversity in medical weblogs can be considered along several dimensions or facets, including *Time, Author, Location, Resource*, *Topic*, *Aspect considered*, *Information Content*, or *Information Type.* An example for values for these dimensions identified for a medical post is given by Figure 1. In this paper, we examine diversity of topic and the diversity of the aspect considered in more detail.

Given a topic T, **topic diversity** concerns the correlation between T and other topics that are frequently used together with T. In this paper, a topic is considered to be a medical concept, in particular a UMLS concept (see section 4.1) dealing with diseases, medical treatments or medications. A topic is highly diverse, if it cooccurs with a large number of other concepts. Even if posts have the same topic, different medical aspects can be considered. While one post 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 post is highly diverse, when its content covers different semantic groups (e.g., symptoms, drugs, procedures).

[...] Most parents are understandably reluctant to start their kids on psychiatric medications. While caution is indicated, I feel such a rigid rejection of their use in children and teenagers is a mistake and may deprive your child of a therapy that could help a lot. Used appropriately and wisely, with ongoing follow-up and psychotherapy, I have seen such medications make a huge positive difference lifting a child's mood and improving the quality of his/her life. [...]

Time: January 2005	**Author:** Dr. J
Resource: Blog	**Topic:** Depression
Aspect considered: Medication	**Information type:** experience
Information content: 50% informative, 50% affective	

Fig. 1. Example for diversity in medical texts

4 Method

We are now presenting a method and measures to study diversity in topic and aspect considered. First, medical content is extracted (see section 4.1) which is then exploited for analysing the diversity (see sections 4.2, 4.3). Since even in a medical weblog, posts can be completely unrelated to health (e.g., dealing with holiday plans or the weather), we exclude posts that are unrelated to medicine and health in a preprocessing step using a classifier based on language models (LingPipe classifier[3]).

4.1 Extracting Medical Content

The medical content of a post is determined by extracting medical concepts, i.e., concepts describing diagnoses, treatments or medications. For this purpose, an existing mapping algorithm (MMTx) is exploited. MMTx, or its java implementation MetaMap [2] is based on natural language processing techniques and maps natural language to concepts of the UMLS Metathesaurus.

The Unified Medical Language System (UMLS[4]) 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**. The concept *atrial fibrillation* belongs for example to the semantic types *Finding and Pathologic Function* that in turn belong to the main group *Disorders*.

In order to determine UMLS concepts for a text, MMTx works in several steps. First, it parses a text into paragraphs, sentences, phrases, lexical elements and tokens. From the resulting phrases a set of (lexical) variants is generated. For the phrases, candidate concepts from the UMLS Metathesaurus are retrieved and evaluated. The best candidates are organized into a final mapping in such a way as to best cover the text. Out of the possible candidates provided by MetaMap, the first proposal out of the highest ranked candidate set is selected as list of extracted concepts which is exploited in our diversity study.

4.2 Determining the Diversity of the Aspect Considered

The diversity of the aspect considered is studied by means of the semantic types and main groups of the concepts extracted from a post and by applying formulas (2) and (3). The formulas determine the proportion of different semantic types (main groups) contained in a text on the overall number of possible types (or groups). A value close to 1 indicates a high diversity while a value close to 0 corresponds to a small diversity.

In addition, we consider the concept diversity $div_{concept}$, i.e. how many different concepts co_d a post contains related to the number of extracted concepts

[3] http://alias-i.com/lingpipe/

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

co (Formula (1)). A value close to zero indicates that the same concepts are used several times, i.e. the diversity of concepts is small. For example, a post can contain only a few different concepts, but these concepts belong to different semantic types and main groups, i.e., it 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. Furthermore, we calculate the 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.

(1) $div_{concept} = \frac{co_d}{co}$ (2) $div_{type} = \frac{types}{135}$ (3) $div_{group} = \frac{groups}{15}$

4.3 Determining the Topic Diversity

For studying diversity of topics, we have to identify relevant topics in the data collection under consideration. We assume that a post deals with a main topic to which in turn a set of subtopics is related. Therefore, the concept representation of texts determined by the method described in section 4.1 is used to (1) determine topics of posts and (2) to identify co-occurring concepts to study topical diversity. For this purpose, we follow a three-step approach:

1. Calculation of concept frequencies per post,
2. Selection of the most frequent concept as ”topic concept”,
3. Identification of relevant concepts related to the topic concept.

Concept frequencies are determined by calculating the number of mentions of a concept in a text. Then, the most frequent concept within the post under consideration is selected. We consider this concept as topic-describing concept (also referred to as ”topic”) for a post. Assuming that the topic of a medical weblog post deals with a disease, a clinical procedure or a medication, a topic concept needs belong to one of the UMLS main groups *Disorder, Procedure* or *Chemicals and Drugs.*

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 of this topic. This results in 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 can be identified within one document collection, this topic is highly diverse. In case a topic-describing concept cooccurs 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.

4.4 Application Example

Having information about different diversity dimensions on hand allows for creating more sophisticated search engines and presentation of results, including grouping search results according to the aspect considered (diagnosis, treatment, medication...). Consider the following scenario: A woman just diagnosed with *breast cancer* enters *breast cancer* into her favourite search engine and receives results grouped into the clusters *disease*, *medications, treatment* and the like. This result structure offers her the opportunity to get a general impression on the different facets of the topic. She can now decide in what kind of information she is interested most. The introduced measures of diversity can be exploited in ranking, e.g. ranking more diverse texts higher.

5 Experiments

5.1 Material

For the analysis in this paper, a set of different medical weblogs written in English and all their posts have been crawled. The resulting data set consists of 5480 posts (patient-written (4343), physician-written (1137)). The weblogs have been selected randomly by collecting addresses of weblogs from the two (medical) weblog search engines Medworm and Medlogs. For comparison reasons we decided to use as an additional data set articles from Yahoo! Encyclopedia[5]. 2777 articles on different topics related to illnesses, treatments and drugs have been collected from this resource. In the following sections, the results are reported when the introduced methods are applied to these data sets.

5.2 Diversity of Aspect Considered

When comparing the concept diversity of the three data sets (see Table 1) it can be seen that the diversity value of the encyclopedia data set is significantly smaller than the one for the weblog datasets. In contrast, the diversity of semantic types and main groups is for the encyclopedia data set much higher. 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. In the weblog datasets the considered aspects are more restricted. Nevertheless, the values for div_{type} are quite small for all three data sets. This shows that from the 135 possible UMLS semantic types only one fourth or one third is covered.

The concept diversity for the categories *Disorders, Procedures* and *Chemicals & Drugs* are similar for all three data sets. In particular for the concept diversity in Procedure concepts, only a small diversity could be ascertained. For the Encyclopedia articles the diversity in disorder-related concepts is higher than the one for blogs. This shows that the spectrum of covered diseases in these articles is higher than in blog posts.

[5] http://health.yahoo.com/ency/

Measure	Patient	Physician	Encyclopedia
$div_{concept}$	0.76	0.70	0.52
div_{type}	0.23	0.27	0.37
div_{group}	0.68	0.78	0.87
$div_{concept}(DISO)$	0.44	0.45	0.53
$div_{concept}(PROC)$	0.18	0.21	0.20
$div_{concept}(CHEM)$	0.30	0.33	0.26

Table 1. Diversity values when considering all semantic types.

5.3 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 topics in physician-written posts only 24 concepts were related. This might be due to the different text length (encyclopedia articles are longer than weblog posts), but is also an indicator to the larger topical diversity in encyclopedia articles.

To study the topical diversity in more detail, we selected the topic *Diabetes Mellitus, Non-Insulin-Dependent.* 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, from the Yahoo! Encyclopedia dataset 259 related topics of 58 different semantic types were extracted and only 60 related concepts of 25 different types were identified in the physician-written posts. We can conclude, that with respect to this topic, the patient-written dataset contains additional information, i.e. covers additional aspects. Figure 2 shows some of the aspects and concepts extracted from the physician-written posts and related to *Diabetes.*

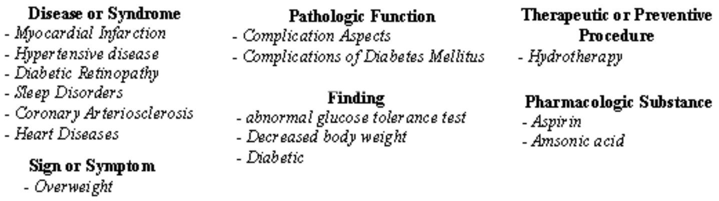

Fig. 2. Related Concepts for the topic "Diabetes" from physician-written posts

6 Conclusion

In this paper, topic and thematic diversity in medical weblogs 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. In order to apply the introduced method to documents of other domains, the underlying domain knowledge has to be replaced or the topics need to be discovered by alternative technologies. 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.

Having information about topic diversity offers the opportunity to get insights into relevant aspects related to a topic that can for example be presented to a user. Furthermore, it can help to improve the retrieval of documents: Documents that consider the same topic but different aspects can be recommend to a user.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *WSDM '09*. ACM.
2. A. Aronson. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. *Proc AMIA Symp*, pages 17–21, 2001.
3. D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *JMLR*, 3, 2003.
4. C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR '08*, pages 659–666, New York, NY, USA, 2008. ACM.
5. M. Cohen. Family medicine meets the blogosphere. *American Academy of Family Physicians. Family Practice Management Web site, http://www.aafp.org/fpm*, 2007.
6. W. Dakka and P. G. Ipeirotis. Automatic extraction of useful facet hierarchies from text databases. In *ICDE '08*, pages 466–475, Washington, DC, USA, 2008. IEEE Computer Society.
7. S. Gollapudi and A. Sharma. An axiomatic approach for result diversification. In *WWW '09*, pages 381–390, New York, NY, USA, 2009. ACM.
8. M. A. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49(4):59–61, 2006.
9. C.-S. Hwang, N. Kuo, and P. Yu. Representative-based diversity retrieval. In *ICICIC '08*, page 155, Washington, DC, USA, 2008. IEEE Computer Society.
10. J. Koren, Y. Zhang, and X. Liu. Personalized interactive faceted search. In *WWW '08*, pages 477–486, 2008.
11. E. Oren, R. Delbru, and S. Decker. Extending faceted navigation for rdf data. In *ISWC 2006*, pages 559–572, 2006.
12. C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW '05*, New York, NY, USA, 2005. ACM Press.