Matching Medical Websites to Medical Guidelines through Clinical Vocabularies in the Context of Website Quality Assessment

D. Rak[1]

[1]1st Faculty of Medicine, Charles University in Prague, Czech Republic

Supervisor: Doc. Ing. Vojtěch Svátek, Dr.

Summary

Modern technology offers a wide array of possibilities to publish almost any content freely on the Internet. Because of the importance and delicacy of medical information, the quality of such texts provided to general public seems to be a serious issue nowadays. Unfortunately, the only feasible way to approve the adequacy of the medical information content is human verification. Best practices in medicine are systematically captured by medical guidelines (MGL), which are provided by renowned medical societies and based on results of Evidence-Based Medicine (EBM).

We propose a simple approach exploiting MGL content as a benchmark for the assessment of a content quality in medical web sites (WS). It is based on the idea that the information content or at least the scope of a medical text is reflected in the domain terminology used. We discuss a possible use of this approach in semiautomatic human-based quality verification and various aspects related to its application.

Concept candidates discovered in a MGL and in the tested web pages are matched to UMLS, yielding sets of used medical terms and corresponding concepts. Several aggregation techniques for MGLs were proposed and tested. The two sets are analyzed for overall similarity at term and concept level.

The method was applied on a selected medical topic employing relevant MGL and 100 WS. All the analyzed web pages fell into five distinct categories (corresponding to the target audience). Aggregations for the MGLs were proposed and tested. The average cosine similarity to MGL across all tested WS reached 0.69 whereas the average similarity calculated per each category varied up to 7,6% against the overall number.

The research done is the first step towards automated evaluation of a medical web page content on the basis of MGLs as the quality standard. We describe further tasks which would improve the outputs of comparison and the possibility of its common application.

Keywords: information quality assessment, clinical vocabularies, unified medical language systems (UMLS), evidence-based medicine (EBM), medical guidelines (MGL), information quality, annotation, similarity, concept, content representation

1. Introduction

Modern technology offers a wide array of possibilities to publish almost any content freely on the Internet. There are many widely available methods of creation and publishing of either static or dynamic web pages today. Although insufficiently, the content is at least somehow linked to the creator or publisher in such classical settings. Besides, there is a variety of new techniques commonly called “Web 2.0“. This technology brings many further possibilities as it allows the readers of the WS to directly contribute and publish their own texts. It encompasses various systems such as blogs, wiki systems, social networks, discussion groups etc. In this case there is in fact no one accountable for the information content except of the system administrator.

Thanks to powerful tools such as Google [1], the lookup of the information on the Internet based on keyword search is even easier than authoring. Search engines constantly scan and index the space of the Internet without any filtering or censorship. The result of user search is returned in the form of a list of pages sorted by their relevance (wrt. the combination of various criteria managed by the search engine provider). Even though providers often boast to provide the user with the 'answers', in fact the engine only returns pages that meet the user search the best. However, the sort criteria completely ignore any content verification or filtering of false information, and they do not distinguish certified web pages (that are assumed to be of a high quality).

The only limitation in this information freedom is just the technical skill of the author of the text. However, the lack of knowledge of the problem area and the competence or qualification to speak about the topic is by no means a limitation. This results in a situation when the user looking for certain information may get many inconsistent answers without having the possibility to distinguish between high-quality information, low-quality information, information influenced by an advertisement, or even intentionally misleading information. Because of the importance and delicacy of medical information this problem is perhaps the most striking in this domain. An easy access to a huge amount of information sources in a varying quality (from meta-analyses to general text) for such an important area of life brings problems in many aspects. Correct information can serve to the user very well and bring him/her many positive effects. In global it can also help achieve many savings in the healthcare system. On the other hand relying on misleading and low-quality data may cause a complete opposite effect. The plausibility of discovered information is thus on the very top position between all the quality measures available.

Another sign of widespread easy access to the great amounts of medical information sources is the information overload. It can concern an ordinary user as well as a medical professional. The result might be the omission of very important information for the case in a pile of other unnecessary data. Often mentioned are the hypothetical problems of complicated communication with a patient previously equipped by wrong information.

Unfortunately, the only feasible way to approve the adequacy of the medical information content is human expert verification today. Experts in the field of WS quality assessment usually evaluate the resources in a complex way. Besides semantics they consider many other rather technical features such as the quality and transparency of presentation, presence of contact information or compliance to the web standards [2]. While there do exist some generic standards for these measures, which might be more or less automatically applied to any kind of web pages including medical ones, presently the only feasible way to approve the adequacy and correctness of the medical information content is manual verification [3]. An example of the system which is designed to support expert decision making in quality assessment is AQUA [4] developed within the frame of the MedIEQ project [5]. Semiautomatic content evaluation would be another important improvement of similar tools allowing better efficiency of an expert work. Once assessed the WS are usually provided with the certificate of quality (e.g. HON [6], WMA [7]) or displayed in specialized portals depending on their quality or a topic category.

The other possible way to ensure the high content quality is the situation when the expert him/herself compiles the text about the topic. Such expert-written texts are often provided by renowned medical societies, which warrants a certain level of quality. Apart from the fact that such practice is very expensive, time consuming and thus in fact unusable in a large scale, the big problem still remains unsolved. Even these high quality texts may still become unrecognized between thousands of other available texts. The main present-day challenges for information science in the area of medical information quality consist in two directions. The first direction consists in the possibility of unambiguous and explicit definition of a unique and consensual version of the truth based on state-of-the-art knowledge. The second challenge is related to the possibility to use this etalon effectively, i.e. find it, compare other documents to it and reference it during the assessment of information quality.

Due to the decentralized creation of new scientific findings, many national specificities occurring in health systems and the existence of a number of organizations aspiring to the position of the highest authority, it is not realistic to expect such a unique and shared version of the truth from any of these entities. The most promising in this context appear to be the activities associated with producing the so-called medical guidelines (MGL) [8]. These documents are systematically prepared and updated by teams of experts and subsequently published under the auspices of prestigious medical societies, medical organizations [9], or agencies specializing in the publication of MGLs [10]. The MGLs are compiled using the principles of Evidence-Based Medicine (EBM) which is based on a hierarchically organized structure of scientific evidence (papers).The aim is to apply primarily the available evidence of the highest strength and significance. The meta-analyses and systematic reviews are on the very top of this hierarchy. MGLs completely cover the area of treatment of the disease in terms of diagnosis, course of the disease, medical procedures, their interchangeability or applicability in different conditions. They even evaluate different methods in relation to their cost or to the difficulties caused to a patient. A very important feature of MGLs is that they are very well structured. Currently, there are already methods aimed to deal with the formalization and with the conversion into an entirely structured electronic version [11]. This can then be implemented for example in hospital systems in combination with electronic healthcare records, or to evaluate information quality of documents on the Internet.

The information quality is defined as the value the information delivers to its user. It implies that a very important role in the information quality is played by its subjectivity. The very quality of information can be viewed from the four different directions (or dimensions). The first group consists of properties directly related to the essence of the text, e.g. accuracy, objectivity and credibility of information. The second dimension features are setting information into the context of other available information (e.g. completeness, timeliness, relevance or value added). The third dimension is related to properties expressing the adoption of text by a reader; therefore it includes properties such as comprehensibility, ease of understanding, conciseness and logical consistency. The last aspect of information quality is associated with the availability of information to users (e.g. ease of obtaining of the information or its updates or security of access). In order to create some information quality assessment framework, the selection of objective characteristics from the options above needs to be performed in the first place. Based on selected options, information quality metrics are to be created.

The subject of this work refers to the objective characteristics of information quality such as completeness of coverage or lack of coverage of the topic, use of professional terminology, accuracy, reliability, verifiability, and accessibility of information. The subjectivity of information is reflected by the authors of texts as they adjust their texts to particular groups of readers. In the field of medical texts on the Internet it is possible to distinguish between texts intended for general public (adult patients or children) and texts for professionals (e.g. physicians and researchers in medicine). Texts targeted for each of these groups differ in many properties falling into the subjective area. For example, the use of accurate medical terminology enhances the accuracy of expression and is usually very appreciated by the professionals. On the other hand, it may significantly reduce the ease of understanding of the text for the non-professional users.

In the group of subjective characteristics, the influence by the reader category is obvious. However, a similar influence of this categorization may be observed even for the characteristics of more objective nature and thus taking it into account during the assessment of information quality seems to be appropriate as well.

2. Objective

The objective of this paper is to propose a simple approach exploiting MGL content as a benchmark for the assessment of an information quality in medical web sites. Clinical vocabularies are used to discover medical terminology in both groups of texts (MGL and WS). Both sets of terminology are then compared based on extracted data.

The WS content quality will be assessed firstly based on general content match (i.e. based on concepts or topics discovered) and secondly based on similarity of the particular terminology used in MGLs. A partial goal is to propose and evaluate suitable methods of aggregation of terminology in MGLs so that only one single standard for WS quality assessment might be applied in the end. The last goal is to evaluate the overall applicability of this approach in the process of semiautomatic quality assessment. The main focus will be placed on the description of strong and weak aspects of the approach and on the evaluation of its impact on the possibility of practical application.

3. Methods

Existing medical guidelines (MGL) and about one hundred web sites (WS) were looked up for the selected medical topic. For both groups were performed searches for medical terminology. The results obtained from the MGLs were used as a benchmark for evaluating the content of WS.

The topic of "lung cancer screening", where the procedure has been tested, was selected based on following criteria. It needed to be clearly identifiable and delimited, there must have been MGLs available for the topic and the subject needs to be reasonably accessible to the general public. The whole work was done on English-language papers (MGL, WS) only. The most suitable MGLs were found by the search in existing databases and catalogues of MGLs. In addition to MGL, the search was carried out for existing WS relevant to this issue. Documents used as a corpus where drawn simply as first one hundred links returned by the search engine Google [1] for the search string "lung cancer screening". These WS were subsequently downloaded by the Scrapbook tool [12] and stored locally. The set of documents was manually rounded to one hundred WS after the removal of broken links or sites that were un-downloadable. Similarly excluded were references to the previously selected MGLs, some of which also ranked in the top 100 results returned by the search engine. These MGLs were excluded to avoid bias in comparison with itself. Based on an estimate of the target group of readers (discussed in the introduction), documents were classified into several categories.

Texts of MGLs were annotated using the tools built over the UMLS Metathesaurus [13], [14]. Initially, mapping was performed using the MMTx (MetaMap Transfer) annotator [15], but the use of the Interactive MetaMap tool [16] proved to be more suitable later. Both tools were developed by the NLM [17], an organization that also develops UMLS. In the first phase, the full texts were processed by either of the tools with the list of terms as the output. For each of these terms corresponding concepts were traced by use of SQL querying over the locally stored UMLS database (containing other data sources such as MeSH, ICD-10, etc.). The result of mapping was a list (or a hierarchical tree) of terms and concepts which served as the set of terminology describing the content of MGLs. Similar mapping to UMLS was performed for all of the WS. Mapping products for the two groups were then preliminarily mutually compared.

Unfortunately, the terminology used in the WS often does not match the terminology used in the MGLs and neither the terminology contained in the UMLS. Even though each concept in the UMLS has assigned a list of synonyms, these terms are again usually scientific terms or names used in other databases of the Metathesaurus. Missing synonyms often comprise colloquial, common, less accurate or abbreviated names of diseases, procedures or medical equipment. These synonyms are necessarily commonly present in texts intended to general public. For instance, the UMLS concept denominated as "Tomography, Spiral Computed" is in reality represented by a range of synonyms, abbreviated or incomplete names and abbreviations such as “CAT Scan, Spiral”, ”Computed Tomography, Spiral”, “Computer-Assisted Tomography, Spiral”, “Computerized Tomography, Spiral”, “CT Scan, Spiral” and the like. This method was very often referred to as only “scan” in the tested WS, which led to the miss with the UMLS terminology or contrary to ambiguous or incorrect mapping. From the perspective of the document content both sets often seemed to differ, even if it was purely syntactic (terminological), rather than semantic difference. The workflow of the method for the final comparison had to be extended so as to be able to take into account even the terms occurring only in WS (i.e. missing in MGLs or in UMLS).

Set of concepts for the selected topic and corresponding synonymic terms (derived from the first round of mapping), were consequently stored outside of UMLS in a different database structure. All the tested WS (or at least few of the WS) previously annotated by the discovered terminology were then manually checked for an overall coverage of the UMLS terms. If not, the missing terms were added one by one to the stored list of terminology. This adjustment was carried out only for the terms clearly classifiable under the chosen concepts (typically there were variants of existing synonyms). This manual step allowed the subsequent more complete mapping of concepts for all the WS and improved their mutual comparison.

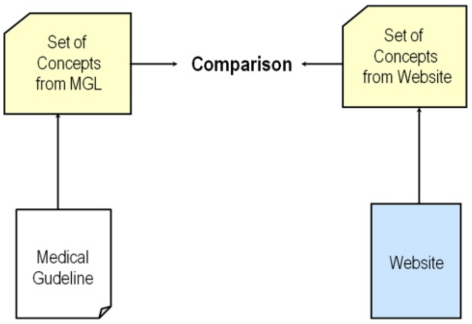


Fig. 1. A diagram showing process of comparison of WS and MGL.

Given that the enriched list of terminology has been stored outside of the UMLS database, the final annotations of WS and MGLs were made using the Super Text Search tool [18], which allows full text searching over the list of documents. Occurring distinct terms and their total number of occurrences in each document were subsequently calculated for annotated files. On the basis of synonymy relations distinct concepts were derived and their occurrences in each of the texts were calculated.

The cosine similarity between the MGLs and WS was calculated from the recorded results. As noted above, there were three suitable MGLs found and used (labeled a0, a1, a2). Because of this, it was necessary to decide, which standard was to be the most appropriate to use for WS comparison. In this context, we proposed four different methods of aggregation of sets of mapped terms or concepts respectively.

The four different aggregations were as follows: "intersection" of sets of terms (labeled a1∩a2∩a0) - i.e. distinct terms that were common to all of the MGLs, then "union" (a1∪a2∪a0) - i.e. distinct terms present in at least one MGL, then "sum" (sum (a1, a2, a0)) - i.e. a simply merged list of all terms found, including the number of occurrences, and finally "weighted sum" (nsum(a1, a2, a0)) - in addition, the merger of sets of documents reflects the scope of each MGL (total number of occurrences for all of the terms). The following table demonstrates the process of construction of the four aggregates on fictive data. Evaluation of the suitability of aggregations was performed using the cosine similarity mutually between all the MGLs and all of the aggregation products. This comparison of mutual similarity was performed first using the full set of all terms found, second based on distinct occurrences of terms present in documents and finally using the mapped concepts.

Comparison of the cosine similarity for all of the WS was made against all the versions of the aggregated standards and against all the MGLs. Comparison was again performed at the level of distinct and total numbers of a medical term which is a number indicating "terminological similarity". Similarly, the similarity was calculated on the basis of distinct (or absolute numbers) of concepts which on the other hand indicates kind of "conceptual similarity". The average cosine similarities of WS against standards were enumerated for each category of documents.

4. Results

The method was applied on a selected medical topic "screening for lung cancer". Test WS were obtained by providing the "lung cancer screening" search string in the Google search engine. Google search returned approximately 2 million records. For instance, similar searches were carried out in the "Yahoo!" search engine [19] (29 million entries) and in the Czech search engine Seznam [20] (only 120 entries). As a corpus of test WS were used the first 100 most relevant results (after adjusting useless links, i.e. broken links and MGLs) returned by Google. Discarded MGLs were positioned in the second half of the top 100 results returned by Google.

All the analyzed web pages fell into five distinct categories (corresponding to the target audience). Aggregations for the MGLs were proposed and tested. The average cosine similarity to MGL across all tested WS reached 0.69 whereas the average similarity calculated per each category varied up to 7.6% against the overall number.

Several MGLs matching the selected topic were found using available information sources, i.e. existing databases or catalogues of MGLs and also freely on the Internet. Three MGLs were chosen and used in the experiment. Discarded MGLs were either nationally or language-specific, they incompletely covered topic or were much more general in the contrary. In one case it was an older version of one of the three used MGLs.

The tested WS were manually classified into several categories depending on their nature and the target group of users of the text. In case of the chosen topic the WS fell into five categories. The first group consists of pages designed for the professionals in medicine, which were represented by 23 % of athe total number of WS. The second important group constituted scientific papers. We have divided this category further into those available in full text or at least in the form of an abstract (23%) and those where there was name or only a very short summary available (5%). An important group of documents were WS targeted for the general public - that is, both for patients (21%) and children (0%). Although the articles for children readers are common for other medical topics (mostly have educational and preventive nature), there were no such documents present in the corpus for the selected topic, probably due to the technical essence of the topic. The last category was created artificially for the texts intended for no particular group of users (28%). This group included namely general reports, statements, newspaper articles and the like.

Tab. 1. Absolute counts of corpus websites in target audience categories.

type	description	count
m	WS for medical professionals	23
p	WS for patients	21
ch	WS for children	0
g	general, news, other	28
mo	scientific papers (full texts or at least abstract)	23
mo/x	scientific papers (restricted access, usually title only)	5

MGLs were annotated by MMTx or the Interactive MetaMap respectively. The results of the process were the texts with mapped scientific terms from UMLS. From the mapped terms it is possible to infer the medical concept which then represent the content of the text. By this procedure 15 distinct concepts relevant to the selected topic area were discovered. Test WS were similarly annotated based on this limited set of medical terminology.

The resulting annotated texts of the two groups (MGLs WS) were then analyzed manually in order to locate gaps in the UMLS mappings. It was found that in MGLs the mapping was almost 100% successful. In contrast, the level of successful mappings in the WS group was estimated to be only about 60%. Both of these findings are fully consistent with expectations and are clearly due to the fact that the UMLS is especially designed to work with texts written by scientific terminology.

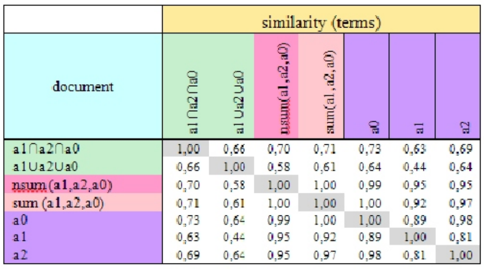
All 15 mapped concepts and all their synonyms contained in the UMLS were saved to a new hierarchical database structure. The original list of synonyms was manually expanded in order to include missing terms identified during the review of annotated WS. Subsequently, all the documents were re-annotated by the extended list of terminology reaching a much higher success rate.

In our case, there were three different MGLs available for the comparison. In order to be able to compare the similarity of WS in the future simply against one single standard, one of the goals was to propose some of aggregation techniques for MGLs. Four different versions of the aggregated sets of terms and concepts were compared using mutual cosine similarity.

As the methods of aggregation where the average similarity shows the best match with the initial MGLs were assessed sum(a1,a2,a0) and nsum(a1,a2,a0). Similarity of these two methods, depending on the method variant ranged from 0.90 to 0.99. The degree of mutual similarity between sets of terms representing the three MGLs was at a similar level (i.e. approaching 1.00). Aggregation based on an intersection and union resulted relatively less suitable based on the mutual similarity analysis. The lowest average similarities were observed for the intersection.

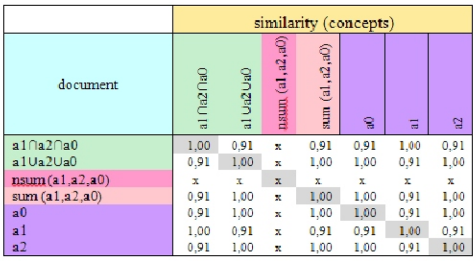
The analysis of similarities was carried out at three levels of the detail. The first method compared all the terms and reflected the number of occurrences in the text as the weight of the term (in the tables referred to as "similarity (terms)"). The second method also worked with the terms, but comparison was limited only to their distinct occurrence in the text ("similarity (distinct terms)"). The last method compared the similarity of concepts mapped through the terms found in the text ("similarity (concepts)"). From this standpoint, the highest average similarity values were achieved on the concept and also on the distinct term level. The lowest average similarity was recorded using the absolute number of occurrences of terms. Summary of results for cross-comparison of MGLs and their aggregations are shown in the following tables.

Tab. 2a. Mutual similarity between sets of terms representing each document. Label a0, a1, a2 corresponds to the three MGLs, "a1∩ a2∩a0" corresponds to the intersection, "a1∪a2∪a0" denotes union, "sum(a1,a2,a0)" denotes sum and "nsum(a1, a2, a0)" denotes the weighted sum. Maximum 100% similarity (i.e. identity) is represented by the value 1. The value of 0 indicates absolute dissimilarity of the two sets.



	similarity (terms)						
document	a1∩a2∩a0	a1∪a2∪a0	nsum(a1,a2,a0)	sum(a1,a2,a0)	a0	a1	a2
a1∩a2∩a0	1,00	0,66	0,70	0,71	0,73	0,63	0,69
a1∪a2∪a0	0,66	1,00	0,58	0,61	0,64	0,44	0,64
nsum (a1,a2,a0)	0,70	0,58	1,00	1,00	0,99	0,95	0,95
sum (a1,a2,a0)	0,71	0,61	1,00	1,00	1,00	0,92	0,97
a0	0,73	0,64	0,99	1,00	1,00	0,89	0,98
a1	0,63	0,44	0,95	0,92	0,89	1,00	0,81
a2	0,69	0,64	0,95	0,97	0,98	0,81	1,00

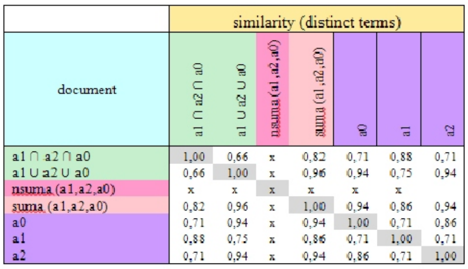
Tab. 2b. Mutual similarity between sets of concepts representing each document. Label a0, a1, a2 corresponds to the three MGLs, "a1∩ a2∩a0" corresponds to the intersection, "a1∪a2∪a0" denotes union, "sum(a1,a2,a0)" denotes sum and "nsum(a1, a2, a0)" denotes the weighted sum. Maximum 100% similarity (i.e. identity) is represented by the value 1. The value of 0 indicates absolute dissimilarity of the two sets.



	similarity (concepts)						
document	a1∩a2∩a0	a1∪a2∪a0	nsum (a1,a2,a0)	sum (a1,a2,a0)	a0	a1	a2
a1∩a2∩a0	1,00	0,91	x	0,91	0,91	1,00	0,91
a1∪a2∪a0	0,91	1,00	x	1,00	1,00	0,91	1,00
nsum (a1,a2,a0)	x	x	x	x	x	x	x
sum (a1,a2,a0)	0,91	1,00	x	1,00	1,00	0,91	1,00
a0	0,91	1,00	x	1,00	1,00	0,91	1,00
a1	1,00	0,91	x	0,91	0,91	1,00	0,91
a2	0,91	1,00	x	1,00	1,00	0,91	1,00

Similarly to the way the sets representing MGLs were mutually compared, the sets representing WS were compared to MGLs too.

Tab. 2b. Mutual similarity between sets of concepts representing each document. Label a0, a1, a2 corresponds to the three MGLs, "a1∩ a2∩a0" corresponds to the intersection, "a1∪a2∪a0" denotes union, "sum(a1,a2,a0)" denotes sum and "nsum(a1, a2, a0)" denotes the weighted sum. Maximum 100% similarity (i.e. identity) is represented by the value 1. The value of 0 indicates absolute dissimilarity of the two sets.



	similarity (distinct terms)						
document	a1 ∩ a2 ∩ a0	a1 ∪ a2 ∪ a0	nsuma(a1,a2,a0)	suma (a1,a2,a0)	a0	a1	a2
a1 ∩ a2 ∩ a0	1,00	0,66	x	0,82	0,71	0,88	0,71
a1 ∪ a2 ∪ a0	0,66	1,00	x	0,96	0,94	0,75	0,94
nsuma (a1,a2,a0)	x	x	x	x	x	x	
suma (a1,a2,a0)	0,82	0,96	x	1,00	0,94	0,86	0,94
a0	0,71	0,94	x	0,94	1,00	0,71	0,86
a1	0,88	0,75	x	0,86	0,71	1,00	0,71
a2	0,71	0,94	x	0,94	0,86	0,71	1,00

This comparison once again took place at three different levels of detail, i.e. at the level of number of terms, at the level of distinct terms and at the level of concepts.

The average similarity across all WS, across all the MGLs (and aggregations) and across all three types of detail reached 0.69. Generally, the lowest similarity was achieved in the analysis at the level of distinct terms (average of 0.56 compared with 0.745 for the concepts and 0.75 for terms). Similarly to mutual comparison of standards, the highest average similarity of corpus WS to MGLs or to their aggregations were found again for the aggregation "sum(a1,a2,a0)", respectively "nsum(a1,a2,a0)". Slightly lower values were found for the non-aggregated MGLs and the lowest value for the "union" and the "intersection" aggregations.

Average category similarities (quantified by each category of documents) deviated from the overall average in the average range of 6.9% for terms and concepts and in the range of 9% for distinct terms. Generally, the highest correspondence of WS and DP was found for the category "mo" i.e. scientific publications (average 0.78) and the lowest for category "g" (general texts) and "mo/x" (incomplete scientific publications) in contrary.

The difference between comparisons using either terms or concepts is not only technical but also rather semantic. When comparing the sets of terms the resulting number describes “similarity of terminology”. The analysis based on the similarity of concepts is actually a “comparison of the content” of both texts. The average similarity of terminology for the corpus of WS reached 0.76 while the average content similarity reached 0.85.

5. Discussion

This work represented the first attempts to compare content of MGLs and WS. Due to this fact we needed to perform a careful selection of the medical topic in order to be able to demonstrate and verify the process of comparison. The topic had to be chosen so that there existed adequate MGLs (i.e. the topic should be completely covered by MGL and on the other hand it should not form only a subset of MGL). For the chosen topic there were available several MGLs in the end. It allowed us to develop and evaluate some potentially useful ways of representing the MGL content in the form of aggregations of sets of terms. This way a single standard for evaluating the content of WS may be created. During the selection of the topic it was also taken into account whether the first 100 WS evenly represented different groups of readers. Of all the possibly expectable groups the corpus of documents did not represented only the group of WS for children.

The possibility to generalize this approach to any medical issue, however, is associated with many complications. The first problem is that the procedure anticipates systematic coverage of the whole domain of medicine by MGLs in the future as it relies on it. Today's practice, however, is far away at least in terms of the coverage and organization of creation of MGLs. MGLs creation is a highly distributed process. MGLs are created irregularly and thematically they cover basically just few of the most important areas. MGLs are also linguistically limited to one particular linguistic area, which constitutes another obstacle to their wider distribution, and in their specific application.

The large influence on the applicability of the method also has the coverage of the medical terminology by the UMLS Metathesaurus. Although the UMLS is regularly updated, expanded to more and more new resources and as the result is has very good coverage of concepts, a number of partial terms in the UMLS is still missing. For each concept it offers a range of synonyms. UMLS can thus be used to map the contained terminology to the words and phrases found in specialized texts. The primary objective of UMLS is to be a dictionary of the correct terminology. For this reason there are many missing terms (particularly colloquial, shortened, incorrect or outdated terms), which results in the fact that the mapping often fails for texts written in an everyday language.

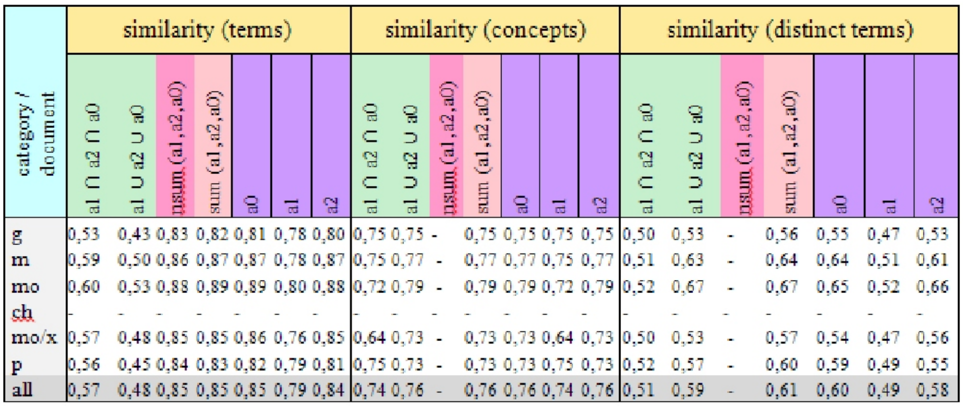
These texts use quite a different terminology from those written in a professional language. This has been also shown in this work. The mapping of the MGLs (written in professional terminology) was almost entirely successful. On the other hand, mapping of WS written in an everyday language achieved success only in 60% of cases.

In order to be able to proceed with the process further and to test the level of the conceptual compliance, we had to extend the list of synonyms manually at one stage. Synonyms were added for all the concepts relating to the selected topic based on discrepancies found in annotated WS. During the manual assessment of WS it proved that the check of the first 10 to 15 papers discloses vast majority of missing terms. The rest of WS were checked just for the sake of completeness. Based on the expanded list of terminology both MGLs and WS were successfully annotated. However, such manual intervention is not generally applicable in bulk for all medical topics and is an obvious weakness of the general application and use of the whole process.

In addition to problems associated with the completeness of the UMLS (i.e. hosting one concept under different names (synonymy)), there are yet other properties of natural languages [21] which pose great obstacles to a reliable term mapping. Probably the most important problem for computer processing of texts is homonymy [22]. For instance: in order to determine precisely which of the meanings of the word is the relevant in a given situation, it is usually necessary to consider the surrounding context and truly understand the meaning of the text.

If we leave aside the problems discussed above, one could imagine using this methodology in a semi-automatic process of assessing the quality of Web documents as follows. The first step in this process would be to establish the topic of the tested text using tools such as Aqua [4]. Identified topic would serve both to search for all MGLs relevant to the topic and also to look up all the relevant terminology using UMLS tools. Subsequently, the evaluation of similarity between text and MGLs would be performed and the resulting degree of similarity could serve as an additional basis for decision-making expert.

Tab. 3. Average cosine similarities of WS categories against MGLg (and aggregations). On the vertical axis there are categories of WS and on the horizontal axis there are MGLs (and aggregations) for the three levels of detail (terms, distinct terms and concepts). Red colour shows maximum values while the blue colour minimum values over the different categories and across all data.



category / document	similarity (terms)							similarity (concepts)							similarity (distinct terms)						
	a1 ∩ a2 ∩ a0	a1 ∪ a2 ∪ a0	nsum (a1,a2,a0)	sum (a1,a2,a0)	a0	a1	a2	a1 ∩ a2 ∩ a0	a1 ∪ a2 ∪ a0	nsum (a1,a2,a0)	sum (a1,a2,a0)	a0	a1	a2	a1 ∩ a2 ∩ a0	a1 ∪ a2 ∪ a0	nsum (a1,a2,a0)	sum (a1,a2,a0)	a0	a1	a2
g	0,53	0,43	0,83	0,82	0,81	0,78	0,80	0,75	0,75	-	0,75	0,75	0,75	0,75	0,50	0,53	-	0,56	0,55	0,47	0,53
m	0,59	0,50	0,86	0,87	0,87	0,78	0,87	0,75	0,77	-	0,77	0,77	0,75	0,77	0,51	0,63	-	0,64	0,64	0,51	0,61
mo	0,60	0,53	0,88	0,89	0,89	0,80	0,88	0,72	0,79	-	0,79	0,79	0,72	0,79	0,52	0,67	-	0,67	0,65	0,52	0,66
ch	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
mo/x	0,57	0,48	0,85	0,85	0,86	0,76	0,85	0,64	0,73	-	0,73	0,73	0,64	0,73	0,50	0,53	-	0,57	0,54	0,47	0,56
p	0,56	0,45	0,84	0,83	0,82	0,79	0,81	0,75	0,73	-	0,73	0,73	0,75	0,73	0,52	0,57	-	0,60	0,59	0,49	0,55
all	0,57	0,48	0,85	0,85	0,85	0,79	0,84	0,74	0,76	-	0,76	0,76	0,74	0,76	0,51	0,59	-	0,61	0,60	0,49	0,58

The document classification into one of the audience categories would help to better interpret the similarity value, or could be considered only for documents belonging to selected categories.

6. Conclusion

The research done is the first step towards automated evaluation of the content of medical web resources using MGLs as a standard of quality. The main goal was to design the process, to evaluate its practical applicability and provide guidelines for further research. At this stage experiments were made on one specific carefully chosen topic for which there existed available appropriate MGLs as well as WS. The topic was elaborated for English-language texts.

In order to obtain a better idea of how to generalize the procedure for any other medical issues, it would be appropriate to make further experiments with randomly selected topics and try to automate the manual steps that process contains. The results show that the similarity between the documents found on the Internet and the MGLs also depends on the category of readers, for whom the text is intended. Category similarities deviated from the overall average in the range of 6.9% (terms and concepts) and 9% (distinct terms) respectively. The highest correspondence of WS and DP was found for scientific publications (average 0.78) while the lowest for category "g" (general texts) and "mo/x" (incomplete scientific publications) in contrary. This categorization of WS was in this work, however, made entirely manually. A promising way to automate this categorization of WS could be the use of the existing functionality of multilingual tool AQUA [2], which was developed for semi-automatic processing of WS. On the other hand, the use of our findings that document classification into a target audience category itself reflects its correspondence with the MGL and thus its quality into some extent may serve directly to expert users of the AQUA system.

Likewise, MGLs search and the selection the best of them was again a purely manual process. In this regard the situation in the future may improve with the way the MGL catalogues on the Internet develop and extend [7], [8]. MGL processing procedures have been designed to enable the automatic aggregation of several available MGLs to one single standard of quality. The best aggregations were selected and tested WS were subsequently compared to this standard.

The next manual step in the procedure was the manual extension of the list of synonyms for UMLS concepts. It was a necessary step for subsequent successful annotation of documents written in an everyday language. On the other hand, it was shown that to find the missing terms it is entirely sufficient to check only the first few annotated WS, ranked according to their search engine relevance.

The set of concepts represents the content of the document only to a certain extent. A great challenge following a simple comparison of sets representing the terminology and content of the documents would be comparing the texts with conflicting claims between the WS and DP [23].

Acknowledgement

Partly supported by MŠMT project no. 1M06014 and by SVV-2010-265513.

References

[1] http://www.google.com (accessed September 10, 2010)

[2] Curro V., Buonuomo P.S., Onesimo R., de RP, Vituzzi A., di Tanna G.L., D'Atri A:. A quality evaluation methodology of health web-pages for non-professionals. Med Inform Internet Med. 2004;29(2):95-107.

[3] Wang Y., Liu Z:. Automatic detecting indicators for quality of health information on the Web. Int J. Med Inform. 2006;May 31.

[4] Stamatakis K., Chandrinos K., Karkaletsis V., Mayer M.A., Gonzales D.V., Labsky D.V., Amigó E., Pöllä M.: AQUA, a system assisting labelling experts assess health web resources. 12th Intern. Symposium for Health Information Management Research (iSHIMR 2007), Sheffield, UK. 2007;18-20 July:75-84.

[5] Mayer M.A., Karkaletsis V., Stamatakis K., Leis A., Villarroel D., Thomeczek C.: "MedIEQ Quality Labelling of Medical Web Content Using Multilingual Information Extraction." Stud Health Technol Inform. 2006;121:183-190.

[6] http://www.hon.ch (accessed September 10, 2010)

[7] Mayer M.A., Leis A., Sarrias R., Ruíz P.: Web Mèdica Acreditada Guidelines: realiability and quality of health information on Spanish-Language websites. In: Engelbrecht R et al. (ed.). Connecting Medical Informatics and Bioinformatics. Proc of MIE2005. 2005;1287-92.

[8] Field M.J., Lohr K.N. (Eds):. Guidelines for clinical practice: from development to use. Institute of Medicine, Washington, D.C: National Academy Press; 1992.

[9] http://www.who.int (accessed September 10, 2010)

[10] http://www.guideline.gov (accessed September 10, 2010)

[11] Vesely A., Zvarová J., Peleska J., Buchtela D., Anger Z.: Medical guidelines presentation and comparing with Electronic Health Record. Int J Med Inform. 2006;Mar-Apr:75(3-4):240-5. Epub 2005; Sep 15.

[12] http://amb.vis.ne.jp/mozilla/scrapbook/ (accessed September 10, 2010)

[13] Lindberg D.A., Humphreys B.L., McCray A.T.: The Unified Medical Language System. Meth Inform Med.1993;32:281-91.

[14] http://www.nlm.nih.gov/research/umls (accessed September 10, 2010)

[15] http://mmtx.nlm.nih.gov (accessed September 10, 2010)

[16] http://skr.nlm.nih.gov/interactive/metamap.shtml (accessed September 10, 2010)

[17] http://www.nlm.nih.gov (accessed September 10, 2010)

[18] http://www.galcott.com/ts.htm (accessed September 10, 2010)

[19] http://www.yahoo.com (accessed September 10, 2010)

[20] http://www.seznam.cz/ (accessed September 10, 2010)

[21] Baud R.H., Ruch P., Gaudinat A., Fabry P., Lovis C., Geissbuhler A.: Coping with the variability of medical terms. Medinfo. 2004;11(Pt 1):322-6.

[22] Rak D., Svatek V., Fidalgo M., Alm O.: Detecting MeSH Keywords and Topics in the Context of Website Quality Assessment. In: The 1st International Workshop on Describing Medical Web Resources (DRMed 2008), held in conjunction with the 21st International Congress of the European Federation for Medical Informatics (MIE 2008), Goteborg, Sweden. 2008: May 27.

[23] Kaiser K., Miksch S.: Versioning Computer-Interpretable Guidelines: Semi-Automatic Modeling of 'Living Guidelines' Using an Information Extraction Method, Artificial Intelligence in Medicine (AIIM), 2008;55-66.

Contact
Mgr. Dušan Rak
U zátiší 545/9
14700 Prague 4– Hodkovičky
Czech Republic
e-mail: dusan.rak@seznam.cz