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

Dusan Rak[1], Vojtech Svatek[2]
[1]1st Faculty of Medicine, Charles University, Prague, Czech Republic
[2]Department of Information and Knowledge Engineering, University of Economics, Prague, Czech Republic

dusan.rak@seznam.cz , svatek@vse.cz

DOI: 10.20470/jsi.v7i1.248

Abstract: *The quality of medical texts provided to general public on the Internet is a serious issue nowadays. Unfortunately the only feasible way to approve the adequacy of the medical information content is human verification today. Best practices in medicine are systematically captured by medical guidelines (MGLs), which are provided by renowned medical societies. We propose a simple approach to exploiting MGL content as 'gold standard' for the assessment of content quality in medical web sites (WS), based on the idea that the information content is reflected in the domain terminology used. Concept candidates discovered in a MGL and in the tested web pages are matched to the UMLS terminological system. In a small case study, MGLs and WSs have been analyzed for similarity at term and concept level. The research is a step towards automated evaluation of WS content on the basis of MGLs as the quality standard.*

Key words: Internet and Web Applications, Information Quality Assessment, Clinical Vocabularies, Unified Medical Language System (UMLS), Medical Guidelines (MGL)

1. Introduction

Modern technology offers a wide array of possibilities to publish almost any content freely on the Internet. There are many available methods of creation and publishing of either static or dynamic web pages today. In such classical settings, the content is somehow (though often loosely) linked to the creator or publisher. 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 web site (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 search tools, the lookup of information on the Internet based on keyword search is even easier than authoring. Search engines constantly scan and index the space of the Internet (mostly) without filtering or censorship. The result of user search is returned in the form of a list of pages sorted by their relevance, which is, in turn, obtained as 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, in the ideal case, meet the user search the best in terms of topic coverage. However the sort criteria completely ignore any content verification or filtering of false information, and they mostly do not recognize certified web pages, which are assumed to have higher quality.

The only limitation in this information publishing freedom is the technical skill of the author of the text. However, the lack of knowledge of the problem area and low competence or qualification to speak about the topic are 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 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. Easy access to a huge amount of information sources in 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 whole healthcare system. On the other hand relying on misleading and low-quality data may cause a completely

opposite effect. Often mentioned by physicians are also communication problems with patients previously equipped by wrong or misinterpreted information from the Internet. The plausibility of discovered information is thus on the very top position between all the quality measures available.

Another implication 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 given case assuming it is buried in a pile of possibly correct but useless information, typically in numerous repliques.

Unfortunately the only widespread way to approve the adequacy of medical information content is human expert verification today, although possibly supported by lightweight information extraction **(Labský, M. & al., 2010)**. Moreover, although experts in the field of WS quality assessment usually evaluate the resources in a complex way, the core of the evaluation still consists in assessing technical features such as quality and transparency of presentation, presence of contact information or compliance to web standards **(Curro, V. & al., 2004)**. 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, reliable automated processing still has to be followed by manual verification **(Wang, Y. & Liu, Z. 2006)**. An example of system designed to support expert decision making in WS quality assessment is AQUA **(Stamatakis, K. & al., 2007)** developed within the frame of the MedIEQ project **(Mayer, M.A. & al., 2006)**. Once assessed, the WS are usually provided with a certificate of quality such as that by Health on the Net[1] or WMA **(Mayer, M.A. & al., 2005)**, or displayed in specialized portals depending on their quality or topic category. To conclude, it is obvious that even lean automatic support targeted on medical content evaluation would be another important improvement of similar tools allowing better efficiency of expert work.

The other possible way to ensure 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 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 can be ranged into two directions. The first direction consists in the possibility of unambiguous and explicit definition of unique and consensual version of the truth based on state-of-the-art knowledge. The second challenge is related to the possibility of using this etalon effectively, i.e. finding it, comparing other documents to it and referencing 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) **(Field, M.J. & Lohr, K.N., 1992)**. These documents are systematically prepared and updated by teams of experts and subsequently published under the auspices of prestigious medical societies, medical organizations such as WHO,[2] or agencies specializing in the publication of MGLs such as the National Guidelines Clearinghouse.[3] 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 primarily apply available evidence that has highest strength and significance. Meta-analyses and systematic reviews are on the very top of this hierarchy. MGLs often completely cover the area of treatment of a given 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 patient. A very important feature of MGLs is that they are well structured. Aside the high culture of writing in their textual versions, there are already methods addressing MGL formalization and their conversion into their entirely structured electronic versions **(Veselý, A. & al., 2005)**. Such computerized MGLs can then be deployed for example in hospital systems in combination with electronic healthcare records, or to evaluate information quality of documents on the Internet.

[1] http://www.hon.ch

[2] http://www.who.int

[3] http://www.guideline.gov

Information quality is typically 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 **(Wang, R.Y. & Strong, D.M., 1996)**. 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 the information into the context of other available information (e.g. completeness, timeliness or 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 the information or its updates or security of access). In order to create any information quality assessment framework, the selection of objective characteristics from the options above needs to be performed in the first place. Based on the 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 of the topic, use of professional terminology, accuracy, reliability, verifiability, and accessibility of information. The subjectivity of information is reflected by the authors 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, similar influence of this categorization may be observed even for characteristics of more objective nature and thus taking it into account during the assessment of information quality seems to be appropriate as well.

2. Website-to-MGL matching approach

The objective of this paper is to propose a simple approach exploiting MGL content as 'gold standard' for the assessment of 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, as outlined in Fig. 1.

The WS content quality is first assessed based on general content match (i.e. on concepts or topics discovered) and then 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 a single standard for WS quality assessment might be applied in the end. A long-term goal is to evaluate the overall applicability of such an approach in the process of semi-automatic quality assessment. Focus is placed on description of strong and weak aspects of the approach and on the evaluation of its possible practical impact.

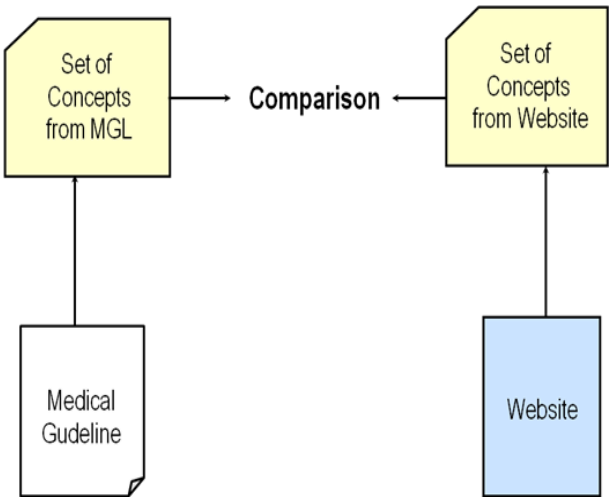

Fig. 1. Outline of the process of comparison of WS and MGL (Source: authors)

2.1 Topic selection and corpus formation

The procedure is deemed suitable for medical topics satisfying the following criteria. First, it needs to be clearly identifiable and delimited, second, there must have been MGLs available for the topic, and last, the subject has to be reasonably accessible to the general public. We focus on English-language documents (MGLs and WS) only. Suitable MGLs are to be found by the search in existing databases and catalogues of MGLs.

The other side of the comparison are the WS to be assessed. Based on a manual estimate of the target group of readers (discussed in the introduction), the WS documents are to be classified into several categories. The required degree of match would be different for each category.

2.2 Term and concept extraction

Texts of MGLs are to be annotated by medical concepts using the tools built over the UMLS[4] Metathesaurus **(Lindberg, D.A. & al., 1993)**. For creating such a mapping we use the MetaMap annotator[5] developed by NLM,[6] the organization that also develops UMLS itself. In the first phase the full texts are processed, yielding a list of terms as output. For each of these terms the corresponding concepts are traced by SQL querying against a locally stored UMLS database (or directly against particular data sources such as MeSH or ICD-10 incorporated within UMLS). The result of mapping is a list (in fact a two-level hierarchical structure) of terms and concepts, which serves as the set of terminology describing the content of MGLs. Similar mapping to UMLS is performed for all of the WS. The mapping products for the two groups are then compared.

The used algorithm of terminology discovery can be generically described as follows (omitting the frequency calculations running in parallel):

Input:
MGL documents: $MGL = \{mgl_1, \ldots, mgl_m\}$
WS documents: $WS = \{ws_1, \ldots, ws_s, \ldots, ws_k\}$
seed set of WS documents: $WS_{seed} = \{ws_1, \ldots, ws_s\} \subset WS$

Processing:
Let T_{init} be the set of all terms $t_i \in (MGL \cup WS)$
Let C be the set of concepts inferred from terms from T_{init} using a mapping tool MT
Let T_{TS} be the set of entry terms corresponding to concepts from C in a terminological system TS
Let T_{man} be the set of terms manually identified in $MGL \cup WS_{seed}$ as corresponding to concepts from C
Let $T_{ann} = T_{TS} \cup T_{man}$ be the annotation term set, and $C_{ann} = C$ be the annotation concept set

Output: T_{ann}, C_{ann}, and the correspondence links between them, denoted as $t_i \rightarrow c_j$.

2.3 Constructing 'gold standard' as aggregation of MGLs

In order to maximize available experimental evidence, we compare the WS with each individual MGL as well as with their aggregations created in multiple different ways. For the purpose of comparison analysis, each of the MGL as well as WS should be represented by a *vector of annotation terms* (for comparison in the term space) or by a *vector of annotation concepts* (for comparison in the concept space). Furthermore, the value of a term/concept can be either *Boolean* or *numerical* (frequency-based; we only assume simple term frequency in our setting). Finally, the *aggregation* of multiple MGL documents can be done in various ways, so as to achieve a 'gold standard' accumulating the medical knowledge from multiple sources.

Let $mgl_1, \ldots, mgl_m$ be the MGL documents, $T_{ann} = t_1, \ldots, t_{maxt}$ the terms from the annotation term set, and $C_{ann} = c_1, \ldots, c_{maxc}$ the concepts from the annotation concept set.

Let $f_i(t_j)$ be the frequency of term t_j in document mgl_i and $f_i(c_j)$ the frequency of concept c_j in document mgl_i, calculated as $\sum f_i(t_k)$; $t_k \rightarrow c_j$. Let the Boolean value for term in document, $b_i(t_j)$, be equal to 1 if

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

[5] http://mmtx.nlm.nih.gov

[6] National Library of Medicine, http://www.nlm.nih.gov

$f_i(t_j)>0$, and 0 otherwise. Analogously, let the Boolean value for concept in document, $b_i(c_j)$, be equal to 1 if $f_i(c_j)>0$, and 0 otherwise.

Straightforward aggregation methods at distinct term level as well as concept level are *intersection*, *union* and *sum*. The weight of term t_j in the aggregated vector can be computed as follows:

$w_{intersection}(t_j) = \min_{i=1..m} b_i(t_j)$ $\quad w_{intersection}(c_j) = \wedge_{i=1..m} b_i(c_j)$

$w_{union}(t_j) = \max_{i=1..m} b_i(t_j)$ $\quad w_{union}(c_j) = \vee_{i=1..m} b_i(c_j)$

$w_{sum\text{-}distinct}(t_j) = \sum_{i=1..m} b_i(t_j)$ $\quad w_{sum\text{-}distinct}(c_j) = \sum_{i=1..m} b_i(c_j)$

For frequencies we consider, at both the term and concept level, a *simple sum*:

$w_{sum}(t_j) = \sum_{i=1..m} f_i(t_j)$ $\quad w_{sum}(c_j) = \sum_{i=1..m} f_i(c_j)$

Finally, we experimented, at the term frequency level, with a *'normalized' sum*:

$w_{nsum}(t_j) = \sum_{i=1..m} f_i(t_j)$. **rlen**(mgl_j)

where **rlen**$(mgl_j)= \sum_{j=1..maxt} f_i(t_j) / \sum_{i=1..m} \sum_{j=1..maxt} f_i(t_j)$, i.e. the relative length of mgl_j in the sense of annotation terms frequency.

2.4 Similarity analysis

Cosine similarity calculation, as the common first choice, is used as the main method for analyzing the data. It was used, first, to identify the relationship of the aggregation products to the original documents, and second, to analyze the similarity between the MGL-based 'gold standard' and the WS corpus.

The evaluation of the aggregations is thus first performed using the cosine similarity mutually between all of the MGLs and all of the aggregation products. The comparison was additionally performed at three different levels of detail – i.e. first using the full set of *all terms* found, second based on *distinct occurrences* of terms present in documents, and last, using the inferred *concepts*.

Comparison of the cosine similarity for all of the WS is done against the aggregations as well as the individual MGLs. The comparison is again performed at three levels of detail. Note that the analysis at the levels of distinct terms and also at the level of total counts of medical terms produces a measure indicating a kind of *terminological similarity* while the similarity calculated on the basis of distinct (or absolute numbers) of concepts on the other hand indicates a kind of *conceptual similarity*. The average cosine similarities of WS against 'gold standard' are also enumerated for each target *audience category* of documents separately.

3. Empirical study

The method was applied on a selected medical topic, "screening for lung cancer". The test WS were obtained by providing the search string "lung cancer screening" to the Google search engine. Google search returned approximately 2 million records. For comparison, similar searches were carried out in the Yahoo! search engine (29 million records) and in the Czech search engine Seznam.cz (only 120 records). As a corpus of test WS we used the first 100 most relevant results returned by Google; some of the previously selected MGLs were however also ranked in the top 100 results and had to be skipped. The four MGLs pages which needed to be discarded were positioned in the second half of the top 100 results returned by Google. The WS were subsequently downloaded by the Scrapbook tool[7] and stored locally. The set of documents was manually rounded to one hundred WS after the removal of broken links or sites that were non-downloadable.

The MGLs corresponding to the selected topic were sought 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. Guidelines labeled 'a1' and 'a2' correspond to the two original guidelines produced by renowned medical societies. Both documents are in the highly formalized NGC[8] format. Besides the key topic-related section 'recommendations' it mainly contains references and other meta-information. Guideline labeled 'a0' is a special guideline synthesis developed by NGC and it directly compares conclusions of guidelines 'a1' and 'a2' and interprets agreements and disagreements

[7] http://www.xuldev.org/scrapbook/

[8] National Guidelines Clearinghouse, http://www.guideline.gov

between the two guidelines. In comparison to 'a1' and 'a2' guidelines, relatively more space is dedicated to the subject topic itself. Unused MGLs were either nationally or language-specific, they incompletely covered the topic (e.g. describing a single detection method) or were much more general on the contrary (i.e. treating cancer in general). In one case it was an obsolete version of one of the three used MGLs.

The tested 100 WS were manually classified depending on their nature and the target group of users of the text. The target audience categorization was borrowed from the EU MedIEQ project **(Mayer, M.A. & al., 2006)**. For the chosen topic the WS fell into five categories, see Tab. 1. (As the total number of corpus WS was 100 (N=100) the numbers in the table also indicate the percentage distribution of WS across categories.). The first group consisted of pages designed for professionals in medicine, and contained 23% of the total number of WS. The second important group was constituted by scientific papers. We divided this category further into papers available in full text or at least as an abstract (23%) and those consisting of title and possibly a very brief summary only (5%). Another group of WS was formed by those targeted for general public: in this case for patients (21%) and children (0%). Although the articles for children readers (mostly of educational and preventive nature) are common for other medical topics, there were no such documents present in the corpus for the selected topic, which was probably due to the technical essence of the topic. The last category was created artificially for texts intended for no particular group of users (28%). This group included, namely, general reports, statements, newspaper articles and the like.

MGLs were annotated using the MetaMap tool. The results of the process were the texts with mapped scientific terms from UMLS. From the mapped terms it was possible to infer medical concepts that represented the content of the texts. By this procedure 13 distinct concepts relevant to the selected topic area were discovered ("Mass Screening", "Tomography, Spiral Computed", "Radiography, Thoracic", "Cytology", "Bronchoscopy", "Breath Tests", "Lung Neoplasms", "Solitary Pulmonary Nodule", "Positron-Emission Tomography", "Biopsy", "Mortality", "lung cancer-associated protein, LCAP", and "Over-diagnosis"). The test WS were similarly annotated based on this filtered set of medical terminology. The resulting annotated texts of the two groups (MGLs and WS) were then analyzed manually in order to locate gaps in the UMLS mappings. It was found that in MGLs the mapping was very successful, reaching 80% of identified medical terms that covered almost 100% of contained concepts. In contrast, the level of successful candidate mappings in the WS group was estimated to be less than 40%. Both these findings were fully consistent with our expectations and were clearly due to the fact that UMLS is primarily designed to work with texts written in scientific terminology.

Tab. 1: Counts of corpus websites for each of the 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

All 13 revealed concepts and their synonyms contained in UMLS were saved to a new two-level hierarchical structure and stored in a database. The original list of UMLS synonyms was manually expanded in order to include missing entry terms (synonyms) from MGLs, which allowed for subsequent almost 100% success in the mapping of MGLs. In order to improve the mapping efficiency for WS, a similar manual review of annotated WS was applied there. The documents from the top of the WS list were examined for missing synonyms. It was observed that the majority of missing terms were identified in the first few documents (in the order originally returned by Google). In the remaining texts the same missing words were just confirmed with only few new entry terms discovered. Because of this fact only the first 10 WS (10%) were used for entry term list expansion, which allowed for the mapping efficiency to rise to estimated 75%. Subsequently, all the documents were re-annotated by the extended list of terminology with much higher success rate.

Unfortunately, the terminology used in the WS often does not match the terminology used in the MGLs and not even 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, i.e. terms that 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 only as “scan” in the tested WS, which led to the missing the UMLS terminology, or on the contrary, to ambiguous or incorrect mapping, depending on the sensitivity used. From the perspective of document content both sets often seemed to differ, even if it was purely due to syntactic (terminological) rather than semantic imperfection of the matching. Because of this, the workflow of the method for the final comparison had to be extended in order to be able to take into account even the terms occurring only in MGLs or only in WS (i.e. missing in UMLS) and to use these synonyms for mapping.

The set of concepts for the selected topic (identified in the first round of mapping) and their corresponding synonymic terms were stored outside of UMLS in a newly-designed database structure. The top 15 WS, previously annotated by the discovered terminology, were manually checked for the overall coverage of the UMLS terms. 17 missing terms found in the first 10 WS (further denoted as ‘seed’ set of WS) were added one by one to the stored list of terminology. This adjustment was carried out only for the terms clearly classifiable under the pre-selected concepts (typically, those were mere variants of existing synonyms). This manual step allowed for subsequent more complete mapping of concepts using the adjusted dictionary for all the WS, thus improving their mutual comparison.

Since the enriched list of terminology was already stored outside the UMLS database, the final annotations of WS and MGLs were made using the Super Text Search tool,[9] which allows full text searching over the list of documents. Distinct terms and their counts for each of the analyzed texts were recorded. On the basis of synonymy and term-concept relations the concepts were derived from the terms and their occurrences, both direct and indirect (i.e. aggregated over corresponding entry terms) for each of the texts were calculated again. The produced sets of terms or concepts, respectively, were subsequently used in the similarity analysis between the vectors representing MGLs and the WS.

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 test the usability of several aggregation techniques. Multiple variants of the aggregated sets of terms and concepts were compared using mutual cosine similarity.

The analysis of similarities among MGLs and their aggregates and similarly between WS and (aggregated and individual) MGLS 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 "term level”). The second method also worked with the terms, but the comparison was limited only to *distinct occurrence of terms* in each of the texts ("distinct term level)”). The last method compared the similarity of *concepts* mapped through the terms found in the text ("concept level)"). The calculations correspond to the weight aggregations described in Section Method; the results for ‘sum-distinct’ at concept level is not included in the tables below due to space limitations (its results did not significantly differ).

The comparison between the results of cross-similarities analyzed at these three levels of detail shows that the highest average similarity values are achieved at the concept level and also at the term level. The lowest average similarity was recorded using distinct terms. The summary of results for cross-comparison of MGLs and their aggregations is shown in Tab. 2. Maximum 100% similarity (i.e. identity) is represented by the value 1. The value of 0 indicates absolute dissimilarity of the two sets.

Similarly to the way the sets representing MGLs were compared mutually, the sets representing WS were compared to MGLs, too. This comparison once again took place at three different levels of detail, i.e. at the level of term frequency, at the level of distinct terms and at the level of concepts, see Tab. 3. The average similarity across all WS, across all the MGLs (and aggregations) and across all three types of detail reached 0.72. Generally, the lowest similarity was achieved in the analysis at the level

[9] http://www.galcott.com/ts.htm

of distinct terms (average of 0.57 compared with 0.78 for the concepts and 0.84 for terms). Similarly to mutual comparison of MGLs, the highest average similarity of corpus WS to MGLs or to their aggregations were found again for the 'sum' and 'nsum' aggregations, respectively. Slightly lower values were found for the non-aggregated MGLs, and the very lowest values for the 'intersection' aggregation.

Tab. 2: Mutual similarity between sets of terms, concepts or distinct terms representing each document. Labels a0, a1, a2 corresponds to the three MGLs.

	term level					concept level						distinct term level					
	nsum	sum	a0	a1	a2	intersection	union	sum	a0	a1	a2	intersection	union	sum-distinct	a0	a1	a2
intersection	0,70	0,71	0,73	0,63	0,69	1,00	0,91	0,84	0,91	1,00	0,91	1,00	0,66	0,82	0,71	0,88	0,71
union	0,58	0,61	0,64	0,44	0,64	0,91	1,00	0,62	1,00	0,91	1,00	0,66	1,00	0,96	0,94	0,75	0,94
nsum	1,00	1,00	0,99	0,95	0,95			1,00									
sum	1,00	1,00	1,00	0,92	0,97	0,91	1,00	1,00	1,00	0,91	1,00	0,82	0,96	1,00	0,94	0,86	0,94
a0	0,99	1,00	1,00	0,89	0,98	0,91	1,00	1,00	1,00	0,91	1,00	0,71	0,94	0,94	1,00	0,71	0,86
a1	0,95	0,92	0,89	1,00	0,81	1,00	0,91	0,93	0,91	1,00	0,91	0,88	0,75	0,86	0,71	1,00	0,71
a2	0,95	0,97	0,98	0,81	1,00	0,91	1,00	0,99	1,00	0,91	1,00	0,71	0,94	0,94	0,86	0,71	1,00

Tab. 3: Average cosine similarities of WS categories against MGLs (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). Maximum values over the different categories and across all data are in italics.

	term level					concept level						distinct term level					
WS category	nsum	sum	a0	a1	a2	intersection	union	sum	a0	a1	a2	intersection	union	sum-distinct	a0	a1	a2
g	0,83	0,82	0,81	0,78	0,80	0,75	0,75	0,87	0,75	0,75	0,75	0,50	0,53	0,56	0,55	0,47	0,53
m	0,86	0,87	0,87	0,78	0,87	0,75	0,77	0,92	0,77	0,75	0,77	0,51	0,63	0,64	0,64	0,51	0,61
mo	0,88	*0,89*	0,89	0,80	0,88	0,72	0,79	0,93	0,79	0,72	0,79	0,52	0,67	*0,67*	0,65	0,52	0,66
mo/x	0,85	0,85	0,86	0,76	0,85	0,64	0,73	*0,93*	0,73	0,64	0,73	0,50	0,53	0,57	0,54	0,47	0,56
p	0,84	0,83	0,82	0,79	0,81	0,75	0,73	0,88	0,73	0,75	0,73	0,52	0,57	0,60	0,59	0,49	0,55
all	**0,85**	***0,85***	**0,85**	**0,79**	**0,84**	**0,74**	**0,76**	***0,90***	**0,76**	**0,74**	**0,76**	**0,51**	**0,59**	***0,61***	**0,60**	**0,49**	**0,58**

Average similarities quantified by each category of documents deviated from the overall average in the average range of 4.1% for terms, 6.7% for concepts and in the range of 9% for distinct terms. Generally, the highest correspondence of WS and MGLs was found for the category 'mo', i.e. scientific publications (average 0.75), and, on the contrary, the lowest similarity was found for the category 'g' (general texts) and 'mo/x' (incomplete scientific publications).

Note that 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 the "similarity of terminology". The analysis based on the similarity of concepts is actually a comparison of the 'content' of both texts. Interestingly, for the corpus of WS with respect to *individual MGLs* the similarity of terminology was higher (between 0.87–0.85) than the content similarity (between 0.74–0.76); the *sum-based aggregation* however yielded the opposite (more expected) result. This phenomenon warrants further investigation.

4. Discussion

This work presented the first attempt to compare the content of MGLs and WS. Due to this fact we needed to perform careful selection of 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 a MGL and on the other hand it should not only form a subset of this MGL). For the chosen topic there were several MGLs available in the end. It allowed us to develop and evaluate some potentially useful ways of representing the MGL content as aggregations of sets of terms. This way a single 'gold standard' for evaluating the content of WS may be created. During the selection of the topic it was also checked that the first 100 WS reasonably represent different groups of intended audience. Of the expectable groups the corpus of documents only lacked the group of WS for children.

The correlation between the similarity of WS to (aggregated) MGLs and the category of these WS is one of the crucial aspects of the experimental approach. The results obtained from this kind of experiment should help tune the parameters to be used in an implemented application for WS quality assessment. For example, the current study corroborates the intuition that WS intended for *patients* might best be compared to the *intersection* of MGLs (as this represents the core notions of the domain, which even patient sites should refer to). On the other hand, the expectation that the *concept* level (reducing the impact of more scholarly terminology often used in the MGLs) should be more relevant for patient WS assessment was not confirmed by the study.

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 basically cover just the most important areas. MGLs are also linguistically limited to one particular language, which constitutes another obstacle to their wider distribution, and in their specific application.

Coverage of medical terminology by the UMLS Metathesaurus also has large influence on the applicability of the method. Although UMLS is regularly updated, expanded to more and more new resources and as a result is has very good coverage of concepts including a range of their synonyms, a number of partial terms in the UMLS is still missing. The primary objective of UMLS is to be a dictionary of 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 everyday language. These texts use quite a different terminology from those written in professional language. This has been also shown in this work: while the mapping of the MGLs (written in professional terminology) was almost entirely successful, the mapping of WS written in everyday language only achieved success in 60% of cases.

In order to be able to proceed with the process further and to test level of the conceptual compliance, we had to extend the list of synonyms manually. Synonyms were added for all the concepts related 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 documents discloses a vast majority of missing terms. The rest of WS were only checked for the sake of completeness. Based on the expanded list of terminology both MGLs and WS have been successfully annotated. However, such manual intervention is not generally applicable in bulk for all medical topics and is an obvious weakness of general application and use of the whole process.

In addition to problems associated with the completeness of UMLS, such as hosting one concept under different names (synonymy), there are also other properties of natural languages **(Baud, R.H. & al., 2004)** that pose great obstacles to reliable term mapping. Probably the most important problem for computer processing of texts is polysemy and homonymy **(Rak, D. & al., 2008)**. In order to precisely determine which of the meanings of the word is the relevant in given situation it is usually necessary to consider the surrounding context and truly understand the meaning of the text.

5. Conclusions

The research presented here 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 outline the overall process, to estimate its practical applicability through empirical exploration, and to 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 general WS. The topic was elaborated for English-language texts. As the feasibility and quality of the evaluation activity necessarily depends on the category of WS and on the aggregation method for multiple MGLs, these two aspects were taken as parameters of the experiment, and multiple settings thereof were tested (though, obviously, by far not exhausting all the possibilities).

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 the described process contains. In particular, the *categorization of WS* with respect to their intended audience was so far made entirely manually, with a risk of introducing a bias. A promising way to automate this categorization could be using the existing functionality of the multilingual tool AQUA **(Stamatakis, K. & al., 2007)**, which was developed for semi-automatic analysis of medical WS. Likewise, the MGLs *search* and the *selection* of the best of them was again purely manual. In this regard the situation may improve in the future as MGL catalogues are being constantly developed and extended on the Internet. The third manual step in the procedure was the *extension of the list of synonyms* for UMLS concepts. It was a necessary step for subsequent successful annotation of documents written in everyday language. On the other hand it was shown that to find the missing terms it is sufficient to only check the first few annotated WS, ranked according to their search engine relevance. Manual processing thus does not present a critical bottleneck at least at the level of prototyping effort. Furthermore, we could leverage on the fact that a steadily growing amount of computerized versions of MGLs will probably contain their own associated machine-readable dictionaries in the future, complementing general nomenclatures, in specific domains.

While the current research focused on principles of experiments potentially yielding the know-how for WS quality assessment, further research will also investigate the usage aspects of a deployed system such as an extended version of AQUA. Such a system would presumably, during WS analysis, produce a warning with a list of MGL-produced concepts (or, terms) not detected in the WS, with additional explanatory information. Manageable size of such a list would have to be assured.

In longer term we plan to pay attention to the fact that a set of concepts only represents the content of the (MGL or WS) document to a certain extent. A further step forward would be representing the terminology and content of the documents in a structured form, so as to detect conflicting claims between the WS and MGL. State-of-the-art methods of MGL text discourse analysis could possibly be adapted to this purpose **(Kaiser, K. & Miksch, S., 2008)**, also leveraging on the logical structure identified for MGLs that underwent structured computerization **(Veselý, A. & al., 2005)**.

Acknowledgements

The research was supported by long-term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

Bibliography

Baud, R.H., Ruch, P., Gaudinat, A., Fabry, P., Lovis, C. & Geissbuhler, A., 2004: Coping with the variability of medical terms. *Medinfo*;11(Pt 1), pp. 322-326

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

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

Kaiser, K. & Miksch, S., 2008: Versioning Computer-Interpretable Guidelines: Semi-Automatic Modeling of 'Living Guidelines' Using an Information Extraction Method, *Artificial Intelligence in Medicine*, 46(1), pp.55-66

Labský, M., Svátek, V. & Nekvasil, M., 2010: Multi-Paradigm and Multi-Lingual Information Extraction as Support for Medical Web Labelling Authorities, *J. Systems Integration*, Vol 1, No 4, pp.3-12

Lindberg, D.A., Humphreys, B.L. & McCray, A.T., 1993: The Unified Medical Language System. *Meth Inform Med.*, 32, pp.281-91

Mayer, M.A., Karkaletsis, V., Stamatakis, K., Leis, A., Villarroel, D. & Thomeczek, C., 2006: MedIEQ – Quality Labelling of Medical Web Content Using Multilingual Information Extraction. *Stud Health Technol Inform.*, 121, pp.183-190

Mayer, M.A., Leis, A., Sarrias, R. & Ruíz, P., 2005: Web Mèdica Acreditada Guidelines: realiability and quality of health information on Spanish-language websites. In: Engelbrecht, R. & al. (ed.). *Connecting Medical Informatics and Bioinformatics. Proc. of MIE2005,* pp. 1287-1292

Rak, D., Svátek, V., Fidalgo, M. & Alm, O., 2008: Detecting MeSH Keywords and Topics in the Context of Website Quality Assessment. In: *1st International Workshop on Describing Medical Web Resources (DRMed 2008)*, Goteborg, Sweden

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

Veselý, A., Zvárová, J., Peleška, J., Buchtela, D. & Anger, Z., 2006: Medical guidelines presentation and comparing with Electronic Health Record. *Int J Med Inform.*, Mar-Apr:75(3-4), pp. 240-245

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

Wang, R.Y.& Strong, D.M., 1996: Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, Spring; 12 (4), pp. 5-34

JEL: L86, **I10**