Recommending medical documents by user profile

Kleanthi Lakiotaki, Angelos Hliaoutakis, Serafim Koutsos and Euripides G.M. Petrakis

Abstract— **The overwhelmed amount of medical information available online, makes the use of automated recommendation methods essential for identifying relevant information according to user profile needs. This paper presents a method to address the problem of medical document classification into documents for medical professionals (experts) and non-professionals (consumers). Documents are represented by terms extracted from AMTEx, a medical document indexing method, specifically designed for the automatic indexing of documents in large medical collections, such as MEDLINE, and then mapped to the UMLS Semantic Network (SN) categories. Multiple Criteria Decision Analysis (MCDA) tools are applied to calculate the membership of each SN category to the document classification. Several factors such as the classification nature of the problem and the incorporation of common readability formulas are also examined.**

I. INTRODUCTION

According to a national survey by the Pew Research Centers Internet & American Life Project in 2013, "*72% of internet users say they looked online for health information of one kind or another within the past year*" [2]. In the world of medical information literature, two major categories of information seekers are mainly identified. The first, often called 'healthcare consumers', or simply consumers, represent those who search in the medical document corpus to find medical information described in simple words, as opposed to 'healthcare experts', that often represent medical professionals.

Over the last few decades, consumer involvement in health care has been significantly increased. Current healthcare consumers are actively involved in seeking health information and based on that information decide about their health. At the same time, the growing health information available on the Internet offers a valuable tool to healthcare consumers.

On the other hand, medical information systems such as MEDLINE[1], are designed to serve health care professional users (expert users in general such as clinical doctors, medical researchers). Typically, expert users are familiar with the type and content of the medical resources (such as the NLM dictionaries and databases) they are using and use medical terminology for their searches.

Based on the above, an automatic system able to characterize medical articles as "consumer specific", or "expert specific" and thus appropriately recommend it, is valuable to both cases, by assisting consumers in managing their personal health information and experts in significantly reducing their effort on information seeking task.

In this work, we investigate on potential improvements to the problem of medical document recommendation by user profile (i.e., consumer users and domain experts) and we show that by incorporating Multiple Criteria Decision Analysis (MCDA) tools, the classification accuracy of existing methods (e.g., Decision Trees) can be further improved. Evaluation results are taken on a subset of MEDLINE documents, the premier bibliographic database of the U.S. National Library of Medicine[2] (NLM). Building upon AMTEx [4] we show that document representations are semantically compact and more efficient, being reduced to a limited number of meaningful multi-word terms (phrases), rather than by large vectors of single-words (as it is typical in classic information systems work) part of which may be void of distinctive content semantics. For this reason, and since medical documents are represented by their content semantics, an unsupervised method like clustering, is unable to separate the documents into two classes based on user profile needs. Moreover, we have already proved [5] that the representation of medical documents by AMTEx terms significantly improves classification. Here, we extend our work by studying more tools (i.e readability formulas) and comparing our results with new datasets.

II. TOOLS & RESOURCES

A. Medical Document Databases

MEDLINE database is a collection of biomedical articles. It consists of medical publications abstracts together with information on the organization of the data, the various data domains, and the relations between them. MEDLINE documents are currently indexed by human experts by assigning to each one, a number (typically 10 to 12) of terms, based on a controlled list of indexing terms, deriving from a subset of the UMLS (Unified Medical Language System) Metathesaurus, the MeSH (Medical Subject Headings) thesaurus.

The **OHSUMED** test collection is a set of 348,566 references from MEDLINE, consisting of titles and/or abstracts from 270 medical journals.

PubMed[3] is a free resource that is developed and maintained by the National Center for Biotechnology Information (NCBI), at the U.S. National Library of Medicine (NLM). It provides free access to MEDLINE, NLM's database of citations and abstracts in the fields of medicine, nursing,

All authors are with the Dept. of Electronic and Computer Engineering, Chania, Crete, Greece klakiotaki@isc.tuc.gr, angelos, euripides@intelligence.tuc.gr, skoutsos@gmail.com

[1]http://www.nlm.nih.gov/bsd/pmresources.html

[2]http://www.nlm.nih.gov/

[3]http://www.nlm.nih.gov/

dentistry, veterinary medicine, health care systems, and preclinical sciences. A strong feature of PubMed is its ability to automatically link to MeSH terms and subheadings.

For consumers, the National Library of Medicine (NLM) provides **PubMed Health**[4]. PubMed Health provides information for consumers and clinicians on prevention and treatment of diseases and conditions. It specializes in reviews of clinical effectiveness research, with easy-to-read summaries for consumers as well as full technical reports. PubMed Health is a service provided by the National Center for Biotechnology Information (NCBI) at the U.S. National Library of Medicine (NLM).

B. Language Processing tools

The **Unified Medical Language System** (UMLS)[5] is a source of medical knowledge developed by the U.S. NLM. UMLS consists of the Metathesaurus, the Semantic Network and the SPECIALIST lexicon.

One of the three components of UMLS, the **Semantic Network**, (SN)[6] is exploited in our experiments. The purpose of the SN is to provide a consistent categorization of all concepts represented in Metathesaurus and a set of useful relationships among these concepts. The Semantic Network contains 133 semantic types and 54 relationships. Every concept in Metathesaurus is assigned to at least one semantic type in the Semantic Network.

The **MeSH** Thesaurus (Medical Subject Headings)[7] is a taxonomy of medical and biological terms and concepts suggested by the U.S NLM. The MeSH terms are organized in IS-A hierarchies, where more general terms, such as "chemicals and drugs", appear in higher levels than more specic terms, such as "aspirin". MeSH (2006) is organized in 15 taxonomies, including 23,884 terms. A term may appear in more than one taxonomy.

WordNet[8] is an on-line lexical reference system developed at Princeton University. WordNet attempts to model the lexical knowledge of a native speaker of English. WordNet v.2.0 (2006) contains around 127,361 terms, organized into taxonomic hierarchies. Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets).

AMTEx[4], implements the C/NC-value [3], a domain-independent method for the extraction of multi-word and nested terms. In this approach, noun phrases are initially selected by linguistic filtering. The subsequent statistical component defines the candidate noun phrase termhood by two measures: C-value and NC-value. More details on AMTEx can be found in [4].

As evidenced by its name, a readability formula can be simply considered as a measure of the ease with which a document can be read. Here, we applied the following readability formulas:

Flesh Reading Ease (Flesh,1948). In the Flesch Reading Ease test, higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. The formula for the Flesch Reading Ease Score (FRES) test is

$$FRES = 206.835 - 1.015(\frac{total\ words}{total\ sentences}) - 84.6(\frac{total\ syllables}{total\ words})$$

Flesch-Kincaid Grade Level. The *Flesch-Kincaid Grade Level Formula* translates the 0-100 score to a U.S. grade level, making it easier for teachers, parents, librarians, and others to judge the readability level of various books and texts. It can also mean the number of years of education generally required to understand this text, relevant when the formula results in a number greater than 10. The grade level is calculated with the following formula:

$$FKG = 0.39(\frac{total\ words}{total\ sentences}) + 11.8(\frac{total\ syllables}{total\ words}) - 15.59$$

The result is a number that corresponds with a grade level. For example, a score of 8.2 would indicate that the text is expected to be understandable by an average student in year 8 in the United Kingdom. Because it is based on adult training manuals rather than school book text, this formula is probably the best one to apply to technical documents.

C. Classification methods

Decision Tree learning is a method for approximating discrete-valued functions,in which the learned function is represented by a decision tree.Learned trees can also be represented as sets of *if-then* rules to improve human readability.These learning methods are among the most popular of inference algorithms and have been successfully applied to a broad range of tasks (diagnose medical cases,credit risk of loan applicants). In our experiments, we applied the J48 classifier, as implemented on WEKA[9] with default parameter values for inducing classification trees.

In Decision Sciences, the field of **Multiple Criteria Decision Analysis (MCDA)** is well established and comes into a large variety of theories, methodologies, and techniques. Here, we implemented the UTA (UTilites Additives) method [6], which exploits special linear programming techniques to infer one or more additive value functions from a weak-order preference structure on a set of alternatives (here medical documents) together with the performances of all the alternatives on all attributes (here SN categories).

III. MEDICAL DOCUMENT RECOMMENDATION BY USER PROFILE

We follow a three phase methodological framework, as described below "Fig. 1A", to prove our assumption that certain Semantic Network category terms are more important that others in medical document classification based on user profile. The following subsections outline the most important steps involved.

[4]http://www.ncbi.nlm.nih.gov/pubmedhealth/

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

[6]http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html

[7]http://www.ncbi.nlm.nih.gov/mesh/

[8]http://wordnet.princeton.edu/

[9]www.cs.waikato.ac.nz/ml/weka

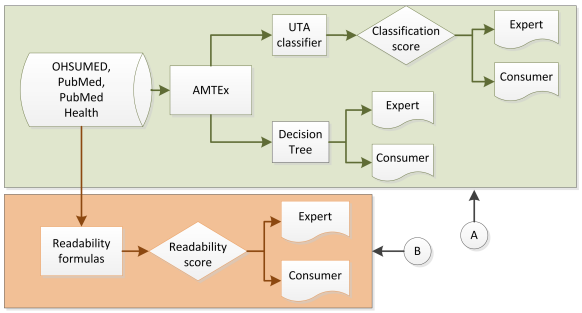

Fig. 1. Classifying medical documents A: by exploiting UMLS Semantic Network. B: based on readability formulas.

A. Medical document collection

As already stated, we used MEDLINE database as a collection of biomedical articles. At first, 474 documents from OHSUMED were classified in both classes of interest (expert and consumer documents), by members of our lab, forming thus the ground truth for our experiments. Also, to test our results in unknown, unexplored medical documents, we created a new dataset, by querying PubMed to retrieve expert and PubMed Health to retrieve consumer documents respectively. The second dataset consisted of 1041 medical documents, 533 expert and 508 consumer documents.

B. Medical document representation

The proposed approach for categorizing MEDLINE documents by user prole relies on the observation that MeSH terms are distinguished into i) *general medical terms* expressing known concepts (e.g., pain,headache) which are easily conceived by all users, ii) *domain specific terms* which are used mainly by experts, iii) *general-non medical terms*. We also assume that the more expert terms a document contains, the higher its probability to be a document for experts. Moreover, since the amount of expert terms in a document is low, even for expert documents, we ignore consumer terms during the modeling process in our experiments and we represent all documents based on expert terms. Consequently, we assume mutual exclusion, meaning that any medical document that is not expert is presumably a consumer document. We combine information from WordNet and MeSH to construct the following three term vocabularies and exploit them to characterize terms as expert or consumer: **Vocabulary of General Terms (VGT)**: these are terms that belong to WordNet but not to MeSH:

$$VGT = (WordNet) - (MeSH)$$

It follows that VGT contains 105.675 general (WordNet) terms. **Vocabulary of Consumer Terms (VCT)**: these are terms that belong to both, WordNet and MeSH:

$$VCT = (WordNet) \cap (MeSH)$$

It follows that VCT contains 7,165 consumer (MeSH) terms. **Vocabulary of Expert Terms (VET)**: these are MeSH terms that do not belong to WordNet:

$$VET = (MeSH) - (WordNet)$$

It follows that VET contains 16,719 expert (MeSH) terms.

Next, documents are represented by terms, extracted by applying AMTEx. In our past work [5], we have already proved that when documents are represented by AMTEx, the categorization performance is significantly increased compared to when the same documents are represented by MMTx[10], the automatic mapping of biomedical documents to UMLS term concepts developed by U.S. National Library of Medicine, or the MeSH method, under which documents are indexed by human experts, based on a controlled list of indexing terms, deriving from a subset of the UMLS Metathesaurus.

Although document contents are summarized by only a few terms, these terms can be any term in the MeSH with almost 24,000 terms, meaning that MCDA should treat any MeSH term as a separate classification criterion which is prohibitive in practice. To ensure in our experiments that all the actions are evaluated on the same basis of criteria, only expert terms are considered in calculating the criteria performances. For example, since we are trying to evaluate the medical documents on their "comprehension difficulty" for a consumer, in other words, on whether they are targeted to experts or not, by considering only expert terms in the criteria performances, we actually adapt a unified measurement scale that reflects the degree of "difficultness" of each attribute. The application of MCDA is enabled by mapping MeSH terms to their more abstract category terms in the Semantic Network of UMLS.

In decision tree analysis, a set of x attributes defines an x-dimensional *description space* in which each instance is a point. Also, since MCDA has mainly focused on the development of comprehensive decision models from small data sets, the total number of expert terms found in a document is too large to be considered as attributes in the decision tree analysis, or criteria in MCDA. Therefore, every term extracted by AMTEx is mapped by a two-layered indexing structure to the UMLS Semantic Network categories, which in turn are considered as criteria.

Only expert terms, as stated above, count in this process and the simple term frequency measure is applied. Hence, a document in the dataset is represented by a 130-dimensional vector (3 of the SN categories do not appear in our corpus) of expert term frequency as:

$$d_i = tf_1, tf_2, ..., tf_n, \rightarrow where\ n = 1, 2, ..., 130$$

and

$$tf_i = \frac{\sum_{j=1}^{k} t_j \rightarrow t_j \in VET}{N}$$

where k is the number of expert terms that belong to the i^{th} SN category and N is the total number of expert terms in d_i (see "Fig.1"). For example, consider that for a document d_i five different expert MeSH terms are extracted by AMTEx, two of which belong to the category "Molecular Function", one to the category "Cell" and the remaining two to "Disease or Syndrom". Then, the value of categories "Molecular

[10]http://ii.nlm.nih.gov/MMTx.shtml

TABLE I

MEAN VALUES OF READABILITY FORMULAS

	OHSUMED		PubMed	
	Consumer	Expert	Consumer	Expert
Flesch Reading Ease	34.31	29.45	40.17	37.03
Flesch Kincaid Grade	13.22	14.45	13.8	12.54

TABLE II

CLASSIFICATION ACCURACY MEASURES-OHSUMED DATA SET

	UTA	Decision Trees
Accuracy (%)	95.80	83.75
Precision (%)	50.85	87.30
Recall (%)	100	83.80

Function" and "Disease or Syndrome" will be 2/5, while of "Cell", 1/5. Therefore, the smaller the number of a category, the less this category contributes to the classification of d_i as expert document. A zero value here means that no expert term from this SN category was extracted. Therefore, by applying the UTA method we attempt to calculate the significance of each category for medical document classification as expert or consumer.

C. Medical document recommendation

To be able to recommend a medical document according to user profile, it must first be labeled as either expert or consumer document. This choice is based upon the utility score calculated based on the final solution that corresponds to the marginal value functions (criteria weights). A linear transformation of the form $\sum_i w_i bi$ where i denotes the SN category and b_i the value of the document under consideration for the specific SN category provides the utility score for every document. The main concern that arises at this point is threshold selection. Youden Index [1] serves here as a global measure of overall diagnostic accuracy and can be used to find an optimal cut-point in discriminating between two classes in ROC curves. This index is be defined as $J = max_i[Sensitivity(i) + Specificity(i) - 1]$ and ranges between 0 and 1. Complete separation of the distributions of the scores for the two discriminated populations results in $J = 1$ whereas complete overlap gives $J = 0$.

IV. RESULTS

As previously mentioned, medical documents cannot be solely distinguished based on their level of difficulty, as this would require a document representation that captures readability level. To further investigate this statement, we computed several different readability formulas. The statistical description of Readability formulas (see Table I for the mean values of the two most characteristic formulas) indicates that the scores for each formula characterize all documents as very difficult to read, independently on the document type (expert or consumer). Moreover, the large overlay of the relative histograms indicated that it is impossible to find a threshold to separate documents according to their readability scores. Next, we apply our proposed approach for recommending medical documents by user profile as this is discussed in Section III, see (Fig.1a). Table II shows prediction accuracy results for the two classification methods (UTA and Decision Trees). Obviously UTA outperforms Decision Trees in all accuracy measures, except precision. Since we evaluate UTA classification ability based on its efficiency in retrieving expert documents, high recall in this case means that UTA retrieves all expert documents in the corpus correctly, however low precision means that also some consumer documents are classified as expert. To further validate our results and study the generalization ability of our classifiers, we used the 1041 PubMed documents explicitly as a test set. In this case, accuracy becomes 70.73% and 63% in the UTA and Decision Trees, respectively.

V. CONCLUSIONS

We investigated the problem of automatic categorization of medical information on two common types of users (consumers and experts) and showed that this problem cannot be solved by simply measuring readability easiness of the documents. Moreover, since medical documents are represented by semantical term vectors, an unsupervised learning method, like clustering, is also insufficient for this problem. In our proposed approach, medical documents were represented by term vectors extracted from AMTEx, a medical document indexing method, specifically designed for the automatic indexing of documents in large medical collections. We proved that the UMLS Semantic Network category terms can act as criteria for the categorization of a medical documents, however their performance play an important role in their classification ability.

VI. ACKNOWLEDGMENTS

Research leading to these results has received funding from the European Communitys Seventh Framework Program (FP7/2007-2013) under grant agreement No 604691 (Project Fi-Star).

REFERENCES

[1] Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal.*, 47(4):458–72, August 2005.

[2] Susannah Fox and Maeve Duggan. Health online 2013. Technical report, Pew Research Centers Internet & American Life Project, 2013.

[3] Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. Automatic recognition of multi-word terms : the C-value / NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[4] Angelos Hliaoutakis, Kaliope Zervanou, and Euripides G.M. Petrakis. The AMTEx approach in the medical document indexing and retrieval application. *Data & Knowledge Engineering*, 68(3):380–392, March 2009.

[5] Kleanthi Lakiotaki, Angelos Hliaoutakis, Serafim Koutsos, and Euripidis G.M. Petrakis. Towards Personalized Medical Document Classification by Leveraging UMLS Semantic Network. In *Lecture Notes in Computer Science*, volume 7798, pages 93–104, 2007.

[6] Yannis Siskos and Evangelos Grigoroudis. UTA methods. In S. Greco J. Figueira and M. Ehrgott, editors, *Multiple criteria decision analysis: State of the art surveys*, pages 297–344. 2005.