Automatic Term Identification by User Profile for Document Categorisation in Medline

Angelos Hliaoutakis and Euripides G.M. Petrakis

Intelligent Systems Laboratory, Electronic and Computer Engineering Dept.
Technical University of Crete (TUC), Chania, Crete, Greece
{angelos,euripides}@intelligence.tuc.gr
http://www.intelligence.tuc.gr

Abstract. We show how term extraction methods such as AMTE$_X$ and MMT$_X$ can be used for the automatic categorisation of medical documents by user profile (novice users and experts). This is achieved by mapping document terms to external lexical resources such as WordNet, and MeSH (the medical thesaurus of NLM).

Key words: term extraction, document indexing, Medline, MeSH, MMTx, AMTEx

1 Introduction

Medical information systems such as MedLine[1] must be capable of providing dedicated, domain specific answers to experts or, simple, easy to comprehend answers to novice users respectively. MedLine documents are currently indexed by human experts by assigning to each one, a number (typically 10 to 12) of terms, deriving from the MeSH[2] (Medical Subject Headings) thesaurus. The automatic mapping of biomedical documents to UMLS Meta-thesaurus[3] term concepts has been undertaken by the U.S. National Library of Medicine (NLM) with the development of MMTx[4] (MetaMap Transfer tool). AMTE$_X$ [1] aims at improving the efficiency of MMT$_X$ based on the extraction and mapping of document terms to the MeSH Thesaurus, rather than the full UMLS Meta-thesaurus mapping of MMT$_X$. It is therefore more selective resulting in more compact representations than MMT$_X$.

In this work, we show how MMT$_X$ and AMTE$_X$ can be used for filtering medical information for targeted audiences such as experts and novice users. An obvious application of this filtering operation will be retrieval on medical information by user profile.

[1] http://www.nlm.nih.gov/databases/databases_medline.html
[2] http://www.nlm.nih.gov/mesh
[3] http://www.nlm.nih.gov/research/umls
[4] http://mmtx.nlm.nih.gov

2 Document Categorisation by User Profile

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. Fig. 1 illustrates the respective categorisation of Medline documents and MeSH terms.

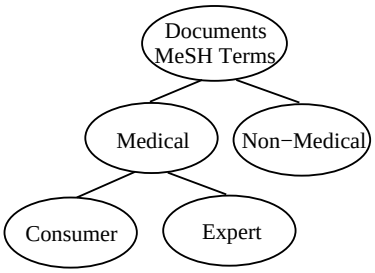

Fig. 1. Categorisation of Medline documents and MeSH terms.

By combining information from WordNet[5] and MeSH the following three term vocabularies are constructed:

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 consumer (MeSH) terms.

The more expert (or consumer) terms a document contains, the higher its probability to be a document suitable for experts (or consumers respectively). Document categorisation by user profile is realized by computing the percentage of expert (VET) and consumer (VCT) terms in a document term vector.

[5] wordnet.princeton.edu

For example, a document with VET% = 0.62 has 62% probability of being a document suitable for experts.

We design an information retrieval method capable of both i) ranking documents by similarity with a query, and ii) bringing documents matching a given user profile higher in the ranked list of similar documents. Documents are represented by term vectors [2] extracted by AMTE_X or MMT_X respectively. As it is typical in information retrieval (IR), the similarity between a query and a document is computed by matching their term vectors according to VSM [2]. More specifically, the query is matched against all Medline documents and the returned list of documents is ranked by decreasing similarity. For ranking query results by user profile we distinguish between the following two cases:

Known user profile: The user identifies her/himself as an expert (or consumer) prior to issuing a query. The similarity score by VSM is multiplied by its percentage of VET (or VCT) terms that is, its probability of being a document for experts (or consumers respectively).

Unknown user profile: The system determines her/his profile from the query. If the query contains at least one expert term, the user is considered to be an expert (a consumer otherwise). Retrievals are then processed similar to the previous case.

3 Evaluation

The experimental results are obtained using OHSUMED, a standard TREC[6] collection of 348,566 medical document abstracts from Medline, published between 1988-1991. OHSUMED is commonly used in benchmark evaluations of IR applications. OHSUMED provides 64 queries and the relevant answer set (documents) for each query. The correct answers were compiled by the editors of OHSUMED and are also available from TREC. For the evaluations, we applied all 64 queries available.

To evaluate our document categorisation method we run a retrieval experiment on the OHSUMED dataset. The results were evaluated against all 64 TREC provided queries and answers; 15 out of the 64 queries contain no expert terms and are suitable for consumer users. The remaining queries are suitable for experts. The objective is to measure the ability of a method in retrieving information for consumer and expert users respectively. We run this experiment twice, once for experts and once for consumer users. A method is deemed successful if it retrieves documents suitable for the particular type of users under consideration.

The candidate methods are i) retrieval using vectors of the manually assigned MeSH terms ii) retrievals using the MMT_X extracted terms and iii) retrievals using terms extracted by AMTE_X. The performance of each method is represented by a precision/recall plot. Each point in such a plot is computed as the average

[6] http://trec.nist.gov/data/t9_filtering.html

precision/recall over all queries. Each method retrieves the best 20 answers for each TREC query so that, each plot contains exactly 20 points.

Fig. 2 (left) illustrates the relative performance of the three retrieval methods examined for the consumer retrieval task. Retrievals with the manually assigned MeSH terms performs better than any other method. This result reveals a tendency of the human indexers to assign simpler terms for the indexed documents. Both AMTE$_X$ and MMT$_X$ perform similarly.

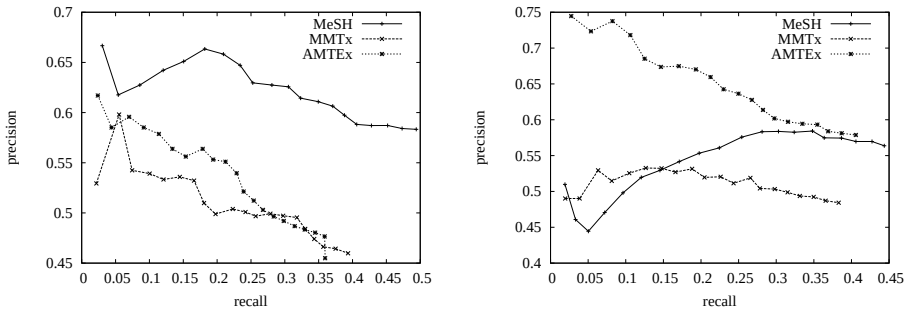

Fig. 2. Average precision/recall for the consumer (left) and expert users (right) retrieval task.

Fig. 2 (right) illustrates results for the retrieval experiment for expert users. AMTE$_X$ outperforms all other methods. This experiment demonstrates the selective ability of AMTE$_X$ towards extracting complex medical terms which can be found in the majority of Medline documents. It also reveals a weakness of manually assigning MeSH terms to documents, as the human indexers may be not familiar with the content complexity and specificity of Medline publications.

4 Conclusions

We introduce to the research community the problem of automatic categorisation of medical publications in Medline by user profile by investigating two common types of users of medical information (i.e., consumers and experts). Based on our experiments, we conclude that AMTE$_X$ selective term output is more effective than MMT$_X$ (the state-of-the-art method of the U.S. NLM).

References

1. Hliaoutakis, A., Zervanou, K., Petrakis, E.: The AMTE$_X$ Approach in the Medical Document Indexing and Retrieval Application. Data and Knowledge Engineering **68**(3) (March 2009) 380–392
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley Longman (1999)