A MeSH Term based Distance Measure for Document Retrieval and Labeling Assistance

Jörg Ontrup*, Tim W. Nattkemper†, Olaf Gerstung, Helge Ritter*
* Neuroinformatics Group, Faculty of Technology, Bielefeld University, Germany
email: jontrup@techfak.uni-bielefeld.de
† Applied Neuroinformatics Group, Faculty of Technology, Bielefeld University, Germany

Abstract—For biomedical and pharmaceutical research, the PUBMED database of the NLM (National Library of Medicine) has become a viable platform. It provides the means for profound investigations of past and related research in daily scientific work. One basic aspect is the search for articles related to a certain research topic. In order to express relatedness many text-mining or document retrieval approaches make use of the “bag of words” model in which unstructured text is represented as a vector of word counts. Since full length articles are not commonly available, many systems generate feature vectors from abstract data only - therefore limiting the explanatory power of their feature space. Since MeSH (Medical Subject Headings) assigned by human experts cover full length articles, we propose for the first time a non-Euclidean document distance measure based on MeSH tree structures. We quantitatively evaluate the approach in comparison to a standard vector space approach and a hybrid version of both. The MeSH-based showed promising results, yet it is still surpassed by the vector space model.

Keywords — MeSH term; information retrieval; data-driven methods

I. INTRODUCTION

The spread of personal computers into all parts of the society and the low cost of memory medias result in an exponential increase of online available data. In the last decade, we observe a strong increase in the number of academic works and commercial tools for document retrieval or text mining in particular. One driving power behind this development is the public interest in the world wide web and its search engines. These allow a fast access to unstructured text data, that may be of potential interest. The drawback of the search engines is, that a search result can consist of a quite large number of hits, that must be manually explored by the user. Additionally, the search is hampered because the user has to enter an adequate boolean expression for an efficient query. Since the investigation of past and related work is one major component in daily scientific work and because of the above reasons, this work becomes more and more time-consuming for the researcher. To support the scientist, several methods were proposed to remedy this situation. Visualization approaches such as ThemeViewTM [12], WebSOM [6] or HSOM [8] provide the user with a “landscape” of documents in which visual attributes are used to indicate relatedness of articles. Once a region of interest is identified, the user can browse that region and retrieve possibly relevant articles. Other systems utilize machine learning algorithms for automated text categorization [5], [7], where each document is tagged with labels from a (usually rather small) set of topics. Another approach aims at the utilization of context information for retrieval of relevant documents [9]. The system shows good results, but it is based on expert knowledge of the whole document domain, thus limiting its application to strongly self-contained document sets.

In the field of biomedical and pharmaceutical research, the MEDLINE database at the National Library of Medicine (NLM) has become the standard platform for scientific investigation. One ground basis for searching the database is the labeling of each item with a set of keywords from the highly standardized MeSH [2] term vocabulary. To guarantee satisfactory search results of constant quality, reproducible labels are an important prerequisite. However, the cost of human indexing of the biomedical literature is high: according to Humphrey [4] it takes one year to train an expert the task of document labeling. Additionally, the labeling process is a highly non-trivial task, which is expressed by a lack of reproducibility. Funk et. al [3] have reported a mean agreement in index terms between 33% - 74% for different experts. Because the improvement of index consistency is such demanding, assistance systems are considered to be a substantial benefit.

Recently, Aronson et al. [1] have presented a highly tuned and sophisticated automated indexing method which yields very promising results. One key element of their approach is the localization of neighbors of a to be labeled document in the database that are most similar to it. To this end they utilize the so called “bag of words” model to represent the document’s abstract and a vector cosine scoring to determine its nearest neighbors. Nevertheless, MeSH terms in the labeled training set contain highly valuable information about the full paper possibly exceeding the information that can be extracted by the bag of words model from the abstract. Therefore, in this paper, we investigate and evaluate an approach to find the neighbors of a document based on a metric derived on a non-Euclidean distance in MeSH tree structures. While first experiments indicate that the alternative MeSH tree metric might provide additional improvements, our results show that the “bag of words” model still seems to be superior when compared to the the pure MeSH tree distance. However a hybrid approach combining both yielded promising results.
II. THE MEDLINE DATABASE

The database contains about 12 million entries, each a scientific article represented at minimum by title, journal title, authors and date of publication covering about 4300 international biomedical journals. For more than 60% of the items, the full abstract is also included. At present the MeSH indexing is done manually by so called skilled subject analysts. Rather than just scanning the title or abstract, the analysts review the complete text of each document and assign descriptors that represent the central concepts as well as every other topic that is discussed to a significant extent [1]. The average number of assigned MeSH terms per document is ten.

To evaluate different approaches for localization of neighboring documents we have selected a training set consisting of 10% (randomly chosen) of all documents in the MEDLINE dating from the 1st quarter of 2002. The test set was determined similarly with articles from the 2nd quarter of 2002. The training and test set contained 7175 and 5496 documents, respectively.

III. METHODS

We compare two approaches to search the MEDLINE database for related articles. The first one follows standard approach of the vector space model for text data (see also [11]). A document \( \alpha \) is mapped to a \( k \)-dimensional feature vector \( v^\alpha \in \mathbb{R}^k \), where the distance between two documents \( d(\alpha, \beta) \) is measured by the dot product of the two vectors \( v_\alpha, v_\beta \), describing the angle between the documents \( \alpha \) and \( \beta \) in the \( k \)-dimensional space. In the second approach we for the first time derive feature vectors for pre-indexed documents, based on the index term positions in the meta-thesaurus MeSH tree. The rational for the MeSH tree distance approach is, that labels of already indexed articles reflect not only the abstract, but the full length paper. Additionally, the highly specific MeSH terms probably suffer from less ambiguity than the more general natural language terms from the abstract data. In the following two subsections we describe the two methods in more detail and present some initial results within our test scenario.

A. The bag of words text model

To construct the vector space for the MEDLINE abstract data, we first build a set of distinct terms \( \{ t_i \} \) for the training corpus of 7175 documents. After word stemming and stop word removal, i.e. neglecting terms with very low and very high frequency, we arrive at a vocabulary of unique word stems \( \{ w_i \} \). For each document \( \alpha \) we then construct a feature vector \( v^\alpha \), where the components \( i \) are determined by the frequency of which word stem \( w_i \) occurs in that document. Following standard practice [10] we choose a \textit{term frequency} \times \textit{inverse document frequency} weighting scheme:

\[
v_i = F_t(w_i, \alpha) \log \left( \frac{N}{F_d(w_i)} \right), \tag{1}\]

where the term frequency \( F_t(w_i, \alpha) \) denotes the number of times the word stem \( w_i \) occurs in document \( d_\alpha \). \( N \) is the number of documents in the training set and \( F_d(w_i) \) the document frequency of \( w_i \), counting the number of documents where \( w_i \) occurs. Distances and therewith dissimilarities of two documents \( \alpha \) and \( \beta \) are computed with the cosine metric

\[
d_{\alpha, \beta} = 1 - \cos(v^\alpha, v^\beta) = 1 - v^\alpha \cdot v^\beta \| v \|. \tag{2}\]

and efficiently implemented by storing the normalized document feature vectors \( v' \).

B. Features from MESH-terms

To construct the second feature space, we consider the list of MeSH index terms \( T(\alpha) = T_1, T_2, \ldots, T_h \) with which the training documents are already labeled. The basic idea of our approach is to represent one document by the subtree, that is built by its MeSH terms and to compute the similarity of two documents by the overlap of their representative subtrees. The similarity values for all pairs \((\alpha, \beta)\) give the distance matrix of the document corpus.

![Fig. 1. From the superposition of all possible subtrees (a), one representative subtree (c) is computed in three steps.](image)

To represent a document by its MeSH term subtree, we face the problem, that many terms are represented more than once in the tree at different positions. For the MeSH tree of 2002 there exist 20603 different MeSH terms with 37864 entries altogether. The term \textit{angioscopy} for example occurs four times, two times under \textit{Diagnosis} and two times under \textit{Surgical Procedures}. The average number of appearance for a MeSH term is \( \approx 1.8 \). Thus, for a document with 12 MeSH indexes the average number of possible subtrees is \( \approx 1.8^{12} = 1157 \). Therefore, in order to represent a document by its MeSH subtree we must select a meaningful representative of the large number of possible subtrees. To this end several heuristics can be applied. In this work, we apply the idea, that the most dense, less ramified subtree is the one that represents the content of the document best. The representative subtree is computed in three steps. First, all possible subtrees are superposed into one tree, where the MeSH terms are located either at a branch or a leaf. In Figure 1 (a) a sample tree is shown for a document that is labeled with three MeSH terms \( T_1, T_2, T_3 \) that all occur two times in the MeSH-thesaurus. Then in a second step for each branch \( i \) a list of those MeSH terms is computed which reside below that branch and \( ii \) the number of MeSH terms in the following subtree is counted. The first and latter are shown as a list in curly brackets and a following number in
Fig. 2: Generation of new MeSH indices involves two different document spaces which are computed from an already labeled training set: For each document \( \alpha \) its bag of words feature vector \( \mathbf{v}^{(\alpha)} \) is computed as illustrated by the squares in region A. Additionally for all document pairs \((\alpha, \beta)\) their pairwise distance \( d_{m}(\alpha, \beta) \) with respect to their MeSH tree similarity is computed as shown by the circles in region B. For a new document \( \gamma \) with a given abstract its bag of words vector \( \mathbf{v}^{\gamma} \) is computed as indicated by the filled square. We can now retrieve the MeSH terms from the \( k_B \) nearest neighbors in space \( A \) or alternatively return the \( k_B \) nearest neighbors from the documents in space \( B \) as visualized by the dashed circles above.

Figure 1 (b), respectively. In a last step for each term one way through the tree is selected. To this end, choose for each term that position in the tree, that is close to the positions of other terms and shares the most branches with other terms. In our example the result is shown in Figure 1 (c).

After computing the unique tree of terms for each document, the distance \( d_m(\alpha, \beta) \) between two documents \((\alpha, \beta)\) is computed as follows. Let \( B_\alpha \) be the set of branches that builds the tree of document \( \alpha \) and \( \Delta(B_\alpha, B_\beta) = (B_\alpha \cup B_\beta) / (B_\alpha \cap B_\beta) \) be the symmetric difference between the two branch sets. Then we define the distance by

\[
d_m(\alpha, \beta) = e^{-|B_\alpha \cap B_\beta|} + \frac{1}{1 + e^{-|B_\alpha \cup B_\beta|}}
\]

The distance is proportional to the complement and anti-proportional to the overlap of the two trees, is symmetric and obeys \( d_m \geq 0 \). Computing the distance for all pairs of documents from one dataset of documents gives the symmetric distance matrix \( D_{\alpha,\beta} \in \mathbb{R}^{+n} \times \mathbb{R}^{+n} \).

To find the \( k_B \) nearest neighbors for a non-labeled document \( \gamma \), we first detect its \( m \) nearest neighbors in the bag of words space described above, respectively their vectors given by the set \( \{ \mathbf{v}^i : |\{ \mathbf{v}^i : d(\mathbf{v}^i, \mathbf{v}^{\gamma}) < d(\mathbf{v}^i, \mathbf{v}^{\gamma}) \} | < m \} \). Because these documents are already labeled, we locate their \( k_B \) nearest neighbors in the tree distance space as indicated in Figure 2, where \( m = 2 \) and \( k_B = 4 \). From these items the pre-indexed terms can be easily extracted.

IV. EXPERIMENTS & RESULTS

In order to create a real world test scenario, we have selected a random set of 10% of all documents in MEDLINE referring to articles published in the 1st quarter of 2002. We parsed the resulting 7175 data items and extracted 251196 different words from their corresponding abstracts. After word stemming and stop word removal, i.e. neglecting all word stems occurring less than 200 and more often than 10000 times, we built a dictionary of 5005 unique word stems for the construction of the vector space model to represent the textual data. The test set was selected from articles published from the 2nd quarter of 2002. Thereby, we modeled the way an actual system would work in the real world: After a certain adaptation time, the system would be applied to new incoming data and use what it had "learned" before. The scheme for generating MeSH terms for new documents using approaches A and B is depicted in Figure 2.

An example of a test document and the corresponding suggestion of MeSH terms resulting from the two approaches is shown in the following two tables. Table I shows an entry with its unique identifier (UI), its title (TI), its abstract (AB) and its expert assigned MeSH terms (MH). Note, that the so-called check tags - high frequency terms in the MeSH hierarchy - were omitted in our example. The system retrieves neighboring documents according to the scheme above and generates a ranked list with respect to frequency of their corresponding MeSH terms. In order to limit the number of retrieved MeSH terms we apply a constant threshold \( \Theta \), such that only a restricted set of MeSH terms is suggested. In our experiments we have chosen \( \Theta = 4 \). Table II shows the system's output in ranked order when applied to the abstract (AB) of Table I. Bold terms indicate a match with the assigned MeSH headings by the human expert. The bracketed numbers following the MeSH terms denote the frequency of the corresponding MeSH heading in the system's suggestion list and can be regarded as a confidence measure.

In order to assess the quality of approach A and B we have computed average precision \( P = \frac{TP}{TP + FP} \) and recall values...
The reliability of frontal and temporal fMRI activations for the determination of hemisphere language dominance was evaluated in comparison with intracarotid amytal testing (IAT). Twenty-two patients were studied by IAT (bilateral in 13, unilateral in 9 patients) and fMRI using a paradigm requiring semantic decisions. [...]

In cases with clear-cut fMRI lateralisation, i.e. consistent lateralisated activation of frontal and temporoparietal language zones, in 9 patients) and fMRI using a paradigm requiring semantic tasks - however at the cost of lower precision. An adjustment of the threshold value \( \Theta \) for \( m = 5 \) would yield results very similar to those for the pure bag of words model. An analysis of the frequency values for approach \( B \) as shown in Table II shows that this method yields rather “concentrated” MeSH headings as the number for the first entries is comparatively high and drops off relatively fast. Future work on the non-Euclidean MeSH tree distance will address other kinds of MeSH tree representations and similarity measures to address this question.

While the hybrid scheme may appear conceptually simple, we think that such an approach might increase the performance of document retrieval in biomedicine and for automated labeling assistance systems, an important goal towards which the present research is meant to be a modest but useful step.

### V. Conclusion & Future Work

Our initial experiments have shown, that the alternative non-Euclidean distance based on MeSH tree structures does not yield results of the same quality as the “good old” bag of words model. Nevertheless, a hybrid approach, where the number \( m \) of neighbours in the vector space used for the lookup in the MeSH tree space is increased does yield slightly better recall values - however at the cost of lower precision. An adjustment of the threshold value \( \Theta \) for \( m = 5 \) would yield results very similar to those for the pure bag of words model. An analysis of the frequency values for approach \( B \) as shown in Table II shows that this method yields rather “concentrated” MeSH headings as the number for the first entries is comparatively high and drops off relatively fast. Future work on the non-Euclidean MeSH tree distance will address other kinds of MeSH tree representations and similarity measures to address this question.

While the hybrid scheme may appear conceptually simple, we think that such an approach might increase the performance of document retrieval in biomedicine and for automated labeling assistance systems, an important goal towards which the present research is meant to be a modest but useful step.

### References


