A k-nearest neighbor based method for improving large scale biomedical document annotation

Khadim Damé
ERIAS, ISPED
University of Bordeaux
F-33000, Bordeaux, FR
Khadim.Drame@u-bordeaux.fr

Fleur Mougin
ERIAS, ISPED
University of Bordeaux
F-330700, Bordeaux, FR
Fleur.Mougin@u-bordeaux.fr

Gayo Diallo
ERIAS, ISPED
University of Bordeaux
F-33000, Bordeaux, FR
Gayo.Diallo@u-bordeaux.fr

Abstract

With the rapid growth of biomedical literature, automated methods for assigning indexing terms to textual documents have received a growing interest. While many efforts have been done towards this direction, it remains a real challenge. Moreover, the issue is even more complicated since full text is not always freely available. In this paper, we propose a k-nearest neighbors (k-NN) based approach which only uses titles and abstracts of a large collection of documents for proposing indexes for their full text. We explore the TF-IDF weighting scheme for document neighbors' retrieval and then investigate several learning methods for annotating documents. Experimental evaluation performed on a standard dataset shows that the proposed method achieves good performance compared with the current state-of-the-art methods, reaching a competitive F-measure of 0.53%.

1 Introduction

The amount of biomedical information is growing rapidly with an abundant production of digital documents (biomedical papers, medical reports, patient discharge summaries, etc.) in the domain. Furthermore, this information is generally expressed in natural language and so in an unstructured form (under textual format), which makes difficult its automated processing. In addition, with their growing volume, effective access to useful information among this large amount of data is necessary. To do so, a suitable representation of the information present in these textual documents is crucial. Controlled vocabularies, such as the Medical Subject Heading (MeSH) thesaurus, are widely used to index biomedical texts (Trieschnigg et al., 2009) and thus to facilitate access to useful information (Díaz-Galiano et al., 2009; Azcárate et al., 2012). The use of these external resources is very important for analyzing biomedical texts and has received a growing interest for improving information retrieval performance. As regards conceptual indexing, concepts defined in thesauri or ontologies are often used to annotate documents. The MeSH thesaurus is a well-known example which is used for indexing Medline citations. Indeed, the latter are indexed manually by the National Library of Medicine (NLM) curators using the MeSH descriptors. Although the task of annotators is now facilitated by a semi-automatic method (Aronson et al., 2004), the rapid growth of biomedical literature makes manual-based indexing approach complex and time-consuming (Huang et al., 2011). Thus, fully automated indexing approaches seem to be essential. While many efforts have been done in this direction, indexing biomedical texts according to specific segments of these texts, such as their title and abstract, remains a real challenge (Tsatsaronis et al., 2012).

In this paper, we propose a k-NN based approach for extracting and selecting relevant MeSH descriptors from the titles and abstracts of a large collection of online scientific articles in order to index their full texts. The principle of the k-NN based approach is to consider the set of concepts (MeSH descriptors, in this case) assigned manually to the k most similar documents (training set) of the document to be indexed. Then, these concepts are ordered by their relevance score so that the more relevant are used to index the document. In a previous work (Huang et al., 2011), authors noted that over 85% of MeSH descriptors relevant for indexing a given document are contained in its 20 nearest neighbors. This seems to better represent

the documents rather than what can be found in their title and abstract solely.

First, we have developed a method to determine similar documents by combining unigram and bigram models with the TF.IDF (term frequency – inverse document frequency) weighting scheme. This latter is used for retrieving the document neighbors. Then, we have investigated different types of features and several learning methods to improve document indexing using the k-nearest neighbors' algorithm. Our work is related to named entity recognition because we identify entry terms of the descriptors in the text. However, our purpose is different in the sense that we use only titles and abstracts of a large collection of documents in order to predict relevant entities for indexing their full text thanks to the exploitation of neighboring documents through a k-NN based approach. These relevant documents could be used later in an Information Retrieval (IR) system. This is a very challenging task, which motivated the recent launch of the BioASQ large-scale biomedical semantic indexing and question answering challenge[1].

The rest of the paper is organized as follows. First, related work concerning biomedical indexing and, more generally, multi-label classification is presented in Section 2. Then, the proposed k-NN based method is described. In Section 4, the experiments are presented while the results are discussed in Section 5. Conclusion and future work are finally drawn in Section 6.

2 Related work

Classical IR methods, which use generally bags of words, present some limitations (Fernández *et al.*, 2011). To address these limitations, knowledge-based approaches, which bring some semantics to the process, are increasingly used. These semantic approaches are often based on resources like thesauri or ontologies (Diallo et al., 2006) and aim at indexing documents with concepts that describe their contents. For indexing textual documents, different approaches have been proposed in the literature: pattern-based methods, machine learning based methods (Jiang et al., 2012) as well as hybrid methods which combine both (Xu et al., 2012).

The pattern-based approaches often rely on linguistic patterns automatically discovered or manually defined to find concepts (or terms) appearing in a document. In particular, the MTI (Aronson et al., 2004) is one of the first attempts to index biomedical documents (Medline articles) using controlled vocabularies. To map biomedical text to concepts from the UMLS Metathesaurus, they used the well-known concept mapper MetaMap and combined its results with the PubMed Related Citations algorithm (Lin and Wilbur, 2007). The combination of these methods results in a list of UMLS concepts which is then filtered and recommended to human experts for indexing citations. Recently, the MTI was extended with various filtering techniques and machine learning algorithms in order to improve its performance. Ruch (2006) has designed a data independent hybrid system for the automatic annotation of medical texts. The first module is based on regular expressions to map texts to concepts while the second is based on a Vector Space Model (VSM) (Salton *et al.*, 1975) considering the vocabulary concepts as documents and documents as queries. Then, the rankers of the two components are merged to produce a final ranked list of concepts with their corresponding weight. The results showed that this method reached good performance, comparable to machine learning based approaches. One limitation of this system is that it may return MeSH terms which match partially the text (Trieschnigg *et al.*, 2009).

The machine learning based approaches, meanwhile, learn a model from a training set constituted of already indexed documents and then use it to classify new documents. Trieschnigg et al. (2009) have presented a comparative study of six systems which aim at classifying medical documents using the MeSH thesaurus. In their experiments, they showed that the k-NN method outperforms the others, including the MTI and the approach developed in (Ruch, 2006). In their work, the k-NN classifier uses a language model to retrieve documents similar to a given document. The relevance of MeSH descriptors is calculated by summing the retrieval scores of documents indexed by these descriptors among the document neighbors. Another k-NN based approach has been proposed in (Huang et al., 2011). A learning-to-rank model is used to compute relevance scores and

[1] http://bioasq.lip6.fr/

[2] Labels are categories used to index documents

then to rank candidate labels[2]. Experiments on two small standard datasets showed that this method achieves better performances than the MTI.

In (Dinh et al., 2013), authors explore a set of factors that affect the effectiveness of biomedical document indexing. They proposed a multi-terminology based concept extraction method using voting techniques. An approximate concept extraction method is used for identifying concepts in documents with their associated relevance scores using each terminology. Then, voting techniques are used to merge lists of concepts obtained with the different terminologies and to select the best concepts. They have shown that their multi-terminology indexing approach improved significantly the performance of IR in the biomedical domain.

On the other hand, indexing biomedical documents where each document of the dataset is assigned one or several categories (also called labels) can be considered as a multi-label classification task. Multi-label classification (MLC) is increasingly studied and especially in text classification (Tsoumakas et al., 2010). Several methods have been developed to deal with this task (Cherman et al., 2011; Spyromitros et al., 2008), which can be categorized into two main approaches (Tsoumakas et al., 2010): the problem transformation approach (Read et al., 2011) and the algorithm adaptation approach (Zhang and Zhou, 2007; Spyromitros et al., 2008; Tsoumakas and Katakis, 2007). The problem transformation approach splits up a multi-label learning problem into a set of single-label classification problems whereas the algorithm adaptation approach adjusts learning algorithms to perform MLC.

In MLC, the k-NN based approach is widely used. This approach was proven efficient for MLC in terms of simplicity, time complexity, computation cost and performance (Spyromitros et al., 2008). Zhang and Zhou (2007) proposed a ML-KNN (for Multi-Label k-NN) method which extends the traditional k-NN algorithm and uses the maximum a posteriori principle to determine relevant labels of an unseen instance. For an instance *t*, the ML-KNN identifies its neighbors and estimates respectively the probabilities that *t* has and has not a label *l* based on the training set, for each label *l*. Then, it combines these probabilities with the number of neighbors of *t* having *l* as a category to compute the confidence score of *l*. Spyromitros *et al.* (2008) propose a similar method, named BR-KNN (for Binary Relevance k-NN), and two extensions of classical MLC. The proposed approach is an adaptation of the k-NN algorithm using Binary Relevance method which trains a binary classifier for each label. Confidence scores for each label are computed using the number of neighbors among the k neighbors that include this label. In addition, an empirical evaluation of these methods was investigated. In (Madjarov et al., 2012), an experimental comparison of several multi-label learning methods is presented. In this work, different approaches were investigated using various evaluation measures and datasets from different application domains. Other recent works address MLC with large number of labels (Bi and Kwok, 2013). Indeed, in many applications, the number of labels used to categorize instances is generally very large. For example, in the biomedical domain, the MeSH thesaurus consisting of thousands descriptors is often used to annotate documents. This large number of descriptors can affect the effectiveness and performance of multi-label models. To address this issue, a label selection based on randomized sampling is performed.

In the following section, we describe our k-NN based approach and experiment the proposed method for indexing biomedical texts.

3 Method

In this work, we propose a biomedical text indexing method which consists of two steps: i) k-nearest neighbors' identification; and ii) document classification.

3.1 K-nearest neighbors' retrieval

For each document, we first retrieve its k-NN from a large dataset annotated previously. For this, we exploit approaches which are based on common words between documents to estimate their similarity (Lin and Wilbur, 2007; Huang *et al.*, 2011). In our work, we determine the similarity degree of a target document with others within the collection according to the cosine measure. This measure is commonly used in text classification and information retrieval within the VSM. The unigrams and bigrams are used in this work to represent the

[2] Labels are categories used to index documents

documents. The principle of the VSM is to represent documents by vectors whose components are the weights of n-grams (unigrams and bigrams) and then to compute the cosine of these weighted vectors to determine the documents' similarity. The TF-IDF weighting scheme is used in this method. Formally, let $C = \{D_1,...,D_n\}$ be a collection of n documents, and $T = \{t_1,...,t_m\}$ the set of the m distinct n-grams occurring in the collection C, a document D_i is thus represented by a multi-dimensional vector :

$V(Di) = \{w_{i1}, \ldots, w_{im}\}$

where w_{ij} is the TF.IDF of the n-gram t_j in the document D_i.

For two given documents D_i and D_j represented respectively by the weighted vectors $V(D_i)$ and $V(D_j)$, their cosine similarity is defined by:

$$Sim(D_i, D_j) = \frac{V(D_i) . V(D_j)}{|V(D_i)||V(D_j)|} \qquad (1)$$

where $V(D_i) \cdot V(D_j)$ is the scalar product of the vectors $V(D_i)$ and $V(D_j)$, and $|V(D_i)|$ and $|V(D_j)|$ their respective norms.

To sum up, our method consists of: i) tokenization and removal of stop words in titles and abstracts; ii) from these preprocessed texts, all the unigrams and bigrams are extracted and stemmed; iii) these extracted n-grams are used with their associated TF-IDF to build the document vectors and thereby to determine the similar documents; iv) for each document, its k most similar documents are recovered.

3.2 Classification of documents

3.2.1 Principle

For a given document, once its k-NNs are retrieved, all categories assigned to these documents are gathered in order to constitute a candidate categories' set likely to index this document. As this can be seen as a classification problem, we have used learning techniques to classify these categories. Simple classifiers are used to determine the categories relevant for indexing a document. The labels are then ranked according to their relevance and the top N most relevant labels are selected, where N is fixed empirically. We have investigated different techniques to determine the optimal value of N. First, we have set N to the number of labels having a relevance score greater than or equal to 0.5. We have also set N to the average size of the labels' sets collected from the k-NN, like in (Spyromitros et al., 2008). Thirdly, we have explored the use of an alternative method to determine the number of labels for each document described in the first BioASQ challenge (Yuqing Mao and Zhiyong Lu, 2013).

To build a classifier, a training set consisting of documents with their associated labels has been constituted. For each document in the training set, its k-NN are retrieved and their associated labels are collected (a list of labels are manually assigned to each document). Each label in this collected set is considered as an instance for the training. Thereafter, this labeled training set is used to build the classifier. We have experimented different classification methods: Naïve Bayes (NB), Decision Trees (DT), Neural Network (NN), and Random Forest (RF) classifiers.

To annotate a given document, the labels collected from its neighbors are represented as the training ones described above and the trained model is then used to estimate the relevance score of each label. Indeed, the model computes, for each label, its probabilities to be relevant and to be irrelevant, and these probability measures are then used to determine its relevance score and therefore to rank candidate labels according to their corresponding scores.

3.2.2 Features extraction

For determining its relevance, each label is represented by a vector of features. In the training step, its class is set to 1 if the label is assigned to the target document or 0 otherwise. In the predicting step, the model uses the label features to estimate this confidence score. Different types of features have been defined based on related work.

For each candidate label, the first feature is the number of neighbors in which it is present. For each candidate label of the target document, the similarity scores of the k nearest documents that are assigned this label are summed and used as a feature. Formally, like in (Spyromitros *et al.*, 2008), let L_j, j=1…n, be the candidate labels' set of a new document D, and D_i, i=1…k its neighbors. The values of features 1 and 2 of the label L_j are respectively:

$$f1(L_j) = \frac{1}{k}\sum_{i=1}^{n} Contain\,(D_i, L_j) \qquad (2)$$

$$f2(L_j) = \frac{1}{k}\sum_{L_j \in D_i} Sim\,(D_i, D) \qquad (3)$$

where the binary function *Contain*(D_i, L_j) outputs 1 if the document D_i contains the label L_j and 0 otherwise; and *Sim*(D_i, D) is the similarity score between D_i and D as described in 3.1.

For each candidate label, we verify whether all its constituent tokens appear in the whole document and consider it as the third feature. This binary feature captures disjoint terms (terms constituted of disjoint words) which are frequent in the medical texts to be processed.

We also computed two other features using term synonyms. Indeed, for indexing the biomedical documents, the MeSH thesaurus is used. The latter is composed of a set of descriptors (called also main headings) organized into a hierarchical structure. Each descriptor includes synonyms and related terms which are known as its entry terms. Hence, for each descriptor, we verify whether one of its entry appears in the document. If this is the case, the fourth binary feature is set to 1 and the descriptor frequency is computed as a value of the fifth feature, otherwise the two features are set to 0.

Finally, another feature is used to verify whether a candidate label is contained in the document title. Our assumption is that if a label is found in the title, this should boost its importance to represent this document.

4 Experiments

In order to evaluate the effectiveness of our method, we have performed two different experiments: one with the official BioASQ challenge dataset provided by the organizers and a derived one from the latter, as described below.

4.1 Datasets

The BioASQ organizers provided a collection of over 4 million documents (constituted by titles and abstracts only) of specific journals in the task 2a of this challenge (Tsatsaronis et al., 2012). It consists of manually annotated articles extracted from PubMed. For the k-NN retrieval, we used a dataset consisting of all articles of this collection published since 2000 (2,268,724 documents). The organizers of the BioASQ challenge then provided sets of PubMed articles not yet annotated which are regarded as test sets to evaluate the participating systems. Participants were asked to classify these test datasets using the MeSH thesaurus. These test documents were then annotated by the PubMed curators for evaluating the results provided by the participating systems.

For the second experiment, from the previous dataset, we extracted all articles published since 2013 (133,770 documents) in which 20,000 randomly selected documents were used for training the classifiers and one thousand as a test set (available in http://lesim.isped.u-bordeaux2.fr/data). The same training set was also used in the first experiment. Like in the training dataset, each document in the test set is assigned a set of labels. These manually assigned labels are used to evaluate our results.

4.2 Evaluation measures

Indexing biomedical documents is considered here as a MLC problem. Instead of one class label, each document is assigned a list of labels. Thus, we used the following measures to evaluate our method: a) example based precision (EBP), b) example based recall (EBR) and c) example based F-measure (EBF) (Tsoumakas et al., 2010). These measures are computed as follows. Let *Yi* be the set of true labels (labels manually assigned to the documents), *Zi* the set of predicted labels and *m* the size of the test set.

$$EBP = \sum_{i=1}^{m} \frac{|Yi \cap Zi|}{|Zi|} \qquad (4)$$

$$EBR = \sum_{i=1}^{m} \frac{|Yi \cap Zi|}{|Yi|} \qquad (5)$$

$$EBF = \sum_{i=1}^{m} \frac{|Yi \cap Zi|}{|Zi| + |Yi|} \qquad (6)$$

4.3 Results

First, we present the results obtained in the task 2a of the BioASQ challenge. For this, we report

results of batch 3 for only the three measures described in section 4.2 (different methods were used for the two first batches but we choose to only present the batch for which our approach obtained the best results). Table 1 shows the results of our method and the one which obtained the highest measures in the different tests of the batch 3. In tests 2 and 5, our best system uses a Naïve Bayes classifier and selects only labels having a confidence score greater than or equal to 0.5 while in the others, the best system sets N to the average size of the neighbors. In most cases, using the average size yielded better or similar results than the others. In the challenge, we did not use the automatic method to fix the number of labels as described in (Yuqing Mao and Zhiyong Lu, 2013) but in the second experiment, this technique has been explored.

Secondly, we evaluated our approach with different configurations in the test set described above and compared the achieved performances. Thus, we tested combinations of different classifiers with different techniques for determining the number of labels for a given document. The evaluation of configurations with the two better classifiers in our experiment, Naive Bayes (NB) and Random Forest (RF), are presented in Tables 2, 3 and 4. The parameter k is empirically set to 25 using cross-validation.

Table 1: Results of our system and the best ones in the different tests of the batch 3. Size is the number of document in the test

Test	Size	System	EBP	EBR	EBF
test 1	2,961	Ours	0.55	0.48	0.49
		Best	0.59	0.62	0.58
test 2	5,612	Ours	0.52	0.50	0.48
		Best	0.62	0.60	0.60
test 3	2,698	Ours	0.55	0.49	0.49
		Best	0.64	0.63	0.62
test 4	2,982	Ours	0.49	0.55	0.49
		Best	0.63	0.62	0.62
test 5	2,697	Ours	0.50	0.53	0.48
		Best	0.64	0.61	0.61

Table 2: Results according to the classifier using 0.5 as the minimal confidence score threshold

Classifier	EBP	EBR	EBF
NB	0.58	0.49	0.49
RF	0.74	0.34	0.43

Table 3: Results according to the classifier using the average size

Classifier	EBP	EBR	EBF
NB	0.51	0.54	0.51
RF	0.52	0.54	0.52

Table 4: Results regarding the classifier using the cut-off method

Classifier	EBP	EBR	EBF
NB	0.56	0.52	0.51
RF	0.61	0.52	0.53

When the minimal score threshold is used, the precision often increases significantly, mainly with the RF classifier but the recall is lower (table 2). Regarding the average size technique, it yields a good recall but the precision decreases slightly (table 3). In this case, the results of both classifiers are similar but the RF one slightly outperforms the NB classifier. The best results are achieved with the cut-off method which balances both precision and recall, and yields the best F-Measure. Except for the minimum threshold technique (table 2) where the NB classifier results are better, the best F-Measure was achieved with the RF classifier (tables 3 and 4). The DT and NN classifiers have been investigated but their results are less interesting. The former yielded worse results while the latter performed very slowly and got results comparable to the RF ones.

5 Discussion

Our experiments show that the proposed approach is promising for medical documents classification. Among the classification systems presented in (Trieschnigg et al., 2009), the k-NN one yielded the best results. When comparing our method with the latter, we use more advanced features to determine the relevance of a candidate label. Indeed, Trieschnigg and his colleagues determine the relevance of a label by summing the retrieval scores of the k neighbor documents that are assigned to the label. In our method, this sum is only considered as one feature among others for determining the confidence scores of labels. While the results of our method did not outperform the MTI system (Mork *et al.*, 2013) which is currently used by the NLM indexers, it got satisfying and comparable results which, need to be improved (0.53 against 0.56 of

F-measure). A direct comparison with the method proposed in (Huang et al., 2011) is not simple since the authors used an old collection different from the BioASQ official datasets which are recent and annotated with the new (2014) MeSH descriptors. As their experiment, when our method is evaluated on 1000 randomly selected documents, it outperforms this method (0.53 against 0.50 for the F-measure). But a comparison with their recent results in the first challenge (Yuqing Mao and Zhiyong Lu, 2013) where they integrate the MTI outputs, their systems performed slightly better than ours (F-measure of 0.55 against 0.53). Compared with two approaches proposed in (Zhu *et al.*, 2013), one based on the MetaMap (Aronson and Lang, 2010) tool and another using IR techniques, our method got better results (0.53 against 0.42 for the F-measure). Our approach outperforms also the hierarchical text categorization approach proposed in (Ribadas-Pena et al., 2013). For our participation in the challenge, the NB classifier was combined to the average size of neighbor's technique to determine relevant descriptors for a given document. In the second experiment, we noted, however, that a combination of RF with the cut-off technique proposed in (Yuqing Mao and Zhiyong Lu, 2013) yielded better results. In addition, we did not use any specific filtering rules like the MTI to improve our performances. This makes our approach generic and its reuse in other domains easier. For the k-NN retrieval, we have investigated the cosine similarity which is widely used in IR. It should be interesting to investigate the combination of this measure with domain knowledge to overcome the limitation of similarity computation based only on common words.

6 Conclusion

In this paper, we presented a k-NN based approach for improving the classification of large collection of biomedical documents. The cosine measure was used with the TF.IDF weighting method to compute similarity between documents and therefore to find the nearest neighbors for a given document. Simple classification methods permitted then to determine the most relevant labels for each document. We have investigated an important feature of the classification problem; the decision boundary which permits to determine the relevant label(s) for a target document. Thus, instead of using voting techniques like in the classical k-NN algorithm, machine learning methods were used to classify documents. Another interesting factor is the parameter k that we have empirically set to 25 using cross-validation. Note that using the RF classifier with the cut-off method yielded the best results in our experiments. We also note that this proposed approach achieved encouraging performance compared with existing methods.

For indexing purpose, the representation of documents as bags of words is limited since similarity between the latter is only based on the words they share. Therefore, we plan to use a wider domain knowledge like the UMLS Metathesaurus in the computation of similarity between documents (exploitation of synonyms and relations) and thus to overcome this limitation. Other features and similarity measures will be studied to improve the performance of our method.

References

Alan Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–236, May.

Alan R Aronson, James G Mork, Clifford W Gay, Susanne M Humphrey, and Willie J Rogers. 2004. The NLM Indexing Initiative's Medical Text Indexer. *Studies in health technology and informatics*, 107(Pt 1):268–272. PMID: 15360816.

Mariano Crespo Azcárate, Jacinto Mata Vázquez, and Manuel Maña López. 2012. Improving image retrieval effectiveness via query expansion using MeSH hierarchical structure. *Journal of the American Medical Informatics Association*:amiajnl–2012–000943, September. PMID: 22952301.

Wei Bi and James Tin-Yau Kwok. 2013. Efficient Multi-label Classification with Many Labels. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28, pages 405–413. JMLR.org.

Everton Alvares Cherman, Maria Carolina Monard, and Jean Metz. 2011. Multi-label Problem Transformation Methods: a Case Study. *CLEI Electron. J.*, 14(1).

Gayo Diallo, Michel Simonet, and Ana Simonet. 2006. An Approach to Automatic Ontology-based Annotation of Biomedical Texts. In *Proceedings of IEA/AIE'06*, Lecture Notes in Artificial Intelligence, vol. 4031, pages 1024–1033, Berlin, Heidelberg. Springer-Verlag.

M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña-López. 2009. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, 39(4):396–403, April. PMID: 19268924.

Duy Dinh, Lynda Tamine, and Fatiha Boubekeur. 2013. Factors Affecting the Effectiveness of Biomedical Document Indexing and Retrieval Based on Terminologies. *Artif. Intell. Med.*, 57(2):155–167, February.

Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. 2011. Semantically Enhanced Information Retrieval: An Ontology-based Approach. *Web Semant.*, 9(4):434–452, December.

Minlie Huang, Aurélie Névéol, and Zhiyong Lu. 2011. Recommending MeSH terms for annotating biomedical articles. *JAMIA*, 18(5):660–667.

Shengyi Jiang, Guansong Pang, Meiling Wu, and Limin Kuang. 2012. An Improved K-nearest-neighbor Algorithm for Text Categorization. *Expert Syst. Appl.*, 39(1):1503–1509, January.

Jimmy Lin and W. John Wilbur. 2007. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8(1):423, October. PMID: 17971238.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Saso Dzeroski. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104.

James G. Mork, Antonio Jimeno-Yepes, and Alan R. Aronson. 2013. The NLM Medical Text Indexer System for Indexing Biomedical Literature. In Axel-Cyrille Ngonga Ngomo and George Paliouras, editors, *BioASQ@CLEF*, volume 1094. CEUR-WS.org.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Francisco J. Ribadas-Pena, Luis M. de Campos Ibañez, Victor Manuel Darriba Bilbao, and Alfonso E. Romero. 2013. Two Hierarchical Text Categorization Approaches for BioASQ Semantic Indexing Challenge. In Axel-Cyrille Ngonga Ngomo and George Paliouras, editors, *Proceedings of the first Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013) , Valencia, Spain, September 27th, 2013*, volume 1094. CEUR-WS.org.

Patrick Ruch. 2006. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658–664, March.

G. Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620, November.

E. Spyromitros, G. Tsoumakas, and Ioannis Vlahavas. 2008. An Empirical Study of Lazy Multilabel Classification Algorithms. In *Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications*, pages 401–406, Berlin, Heidelberg. Springer-Verlag.

Dolf Trieschnigg, Piotr Pezik, Vivian Lee, Franciska de Jong, Wessel Kraaij, and Dietrich Rebholz-Schuhmann. 2009. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics (Oxford, England)*, 25(11):1412–1418, June. PMID: 19376821.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Éric Gaussier, Patrick Gallinari, Thierry Artières, Michael R. Alvers, Matthias Zschunke, and Axel-Cyrille Ngonga Ngomo. 2012. BioASQ: A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In *Information Retrieval and Knowledge Discovery in Biomedical Text, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012*, volume FS-12–05. AAAI.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-Label Classification: An Overview. *IJDWM*, 3(3):1–13.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis P. Vlahavas. 2010. Mining Multi-label Data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer.

Yan Xu, Kai Hong, Junichi Tsujii, and Eric I.-Chao Chang. 2012. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *JAMIA*, 19(5):824–832.

Yuqing Mao and Zhiyong Lu. 2013. Ncbi at the 2013 bioasq challenge task: Learning to rank for automatic mesh indexing. Technical report.

Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048.

Dongqing Zhu, Dingcheng Li, Ben Carterette, and Hongfang Liu. 2013. An Incremental Approach for MEDLINE MeSH Indexing. In Axel-Cyrille Ngonga Ngomo and George Paliouras, editors, *Proceedings of the first Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013 (CLEF 2013) , Valencia, Spain, September 27th, 2013*, volume 1094. CEUR-WS.org.