Automatic MeSH Term Assignment and Quality Assessment

Won Kim†, PhD, Alan R. Aronson*, PhD, and W. John Wilbur†, MD, PhD

†National Center for Biotechnology Information (NCBI)
***Lister Hill National Center for Biomedical Communications (LHNCBC)**
National Library of Medicine, Bethesda, MD 20894

For computational purposes documents or other objects are most often represented by a collection of individual attributes that may be strings or numbers. Such attributes are often called features and success in solving a given problem can depend critically on the nature of the features selected to represent documents. Feature selection has received considerable attention in the machine learning literature. In the area of document retrieval we refer to feature selection as indexing. Indexing has not traditionally been evaluated by the same methods used in machine learning feature selection. Here we show how indexing quality may be evaluated in a machine learning setting and apply this methodology to results of the Indexing Initiative at the National Library of Medicine.

INTRODUCTION

One of the important problems frequently solved by machine learning methods is the problem of document classification. Generally one has a reasonably large set of documents that have already been classified (training set) and one seeks to learn from this given data how to classify an additional set of documents. A number of machine learning methods are available to solve this problem including Naïve Bayes, k-nearest neighbors, decision trees, neural networks, and support vector machines to name the most common[1-3]. For all of these methods and others an important issue is how the documents are represented by features (most often words or phrases from the text of the documents). It has proven helpful to be selective in choosing the features to represent the documents and a number of measures have been developed which provide some guidance to the quality of the individual terms to be selected. These measures generally reflect how strongly a feature's occurrence correlates with the known classification on the training set of documents[3-6].

The problem of document retrieval in answer to a query faces the same issue of feature selection that is faced in machine learning; however, the same methods cannot be applied because one does not have a training set on which to base feature selection. In fact one seeks to classify the documents into two classes, the relevant and the nonrelevant, and for each query that may be posed the classification will generally be different and unknown at the time of indexing. In this setting we seek methods of improved indexing of individual documents and methods of evaluating such indexing.

Automatic indexing generally begins with the single words that are in a text minus stop words. Though efforts have been made to discard more than stop words, no generally effective method of improvement has been found[7]. Attempts have also been made to add phrases from the text to single words, but again this has not proved generally useful[8-10]. For the MEDLINE® database it has been observed that adding the MeSH® terms to the text does give an improvement in performance[11, 12]. This suggests a strategy of automatic selection of MeSH terms for addition to documents and raises the question as to whether such indexing could at least partially replace the humanly selected MeSH terms that are currently being added to MEDLINE documents. Might one see the same improvement in document retrieval from such automatically selected MeSH terms?

The Indexing Initiative System (IIS) at the NLM provides a number of methods for automatically computing MeSH terms that could be added to a document prior to standard MeSH indexing. Here we examine the best such strategy[21] and ask how it compares with the humanly assigned MeSH terms for purposes of document retrieval. Instead of using some available search engine to make the comparison we here introduce a method we refer to as undertrained Bayesian weighting. This is a form of Naïve Bayesian machine learning that can be applied to test sets and gives retrieval results generally superior to vector methods (and independent of them). Our results show that IIS assigned MeSH terms perform on a par with humanly assigned MeSH terms on the three test sets we consider.

TESTING METHODOLOGY

Test Sets. In this study we use three sets of queries and documents with human judgments of relatedness.

1. OHSUMED collection. This database consists of 348,566 MEDLINE documents from 270 medical journals over a five-year period (1987-1991) and 105 queries generated by novice physicians[11].

2. Small MEDLINE document collection produced by Haynes and modified by Hersh. This set consists of 75 queries and 2,344 MEDLINE documents for which both the title and abstract are present[13, 14]. We will refer to this set as 2,344 MED.
3. MEDLINE document Test Set B constructed by Wilbur. This set consists of 100 query MEDLINE documents and for each query document 50 lexically close MEDLINE documents. The resultant 5,000 query-document pairs have been judged by a panel of seven judges[15, 16].

Test Set Indexing. Preprocessing of text consists of three steps: i) all stop words and punctuation marks are removed; ii) all alphabetic characters are lowercased; iii) all non-alphanumeric characters are replaced by blanks. No stemming is done. The resulting strings demarcated by white space minus any strings that are purely numeric are the index terms derived from the text. This procedure is applied uniformly to titles, abstracts, and to MeSH terms also when they are included in the indexing. All terms derived from different sources are used to make a single index.

Undertrained Bayesian retrieval. Given a partition of the database into two sets, those relevant to a query *q* and those non-relevant, Naïve Bayesian machine learning provides a method of producing term weights that are ideal for reproducing this classification on new material[5, 17-19]. If such weights are applied to the database from which they were derived they are overtrained Bayesian weights (OBW). However, we can perform a form of crossvalidation by removing a document *d* from the database, deriving the Bayesian weights from the classification of the remaining documents and then applying those weights to *d*. Then *d*'s classification has not contributed to the derivation of the weights that are used to score *d* and in this case we have undertrained Bayesian weights (UBW). Because the UBW are close to ideal we use them in our evaluation of the quality of indexing.

Local weighting. Most models of document retrieval perform best with local weighting of terms that is dependent on the frequency of the terms in the documents. We use a local weight to combine with the global UBW. If f_{td} denotes the frequency of term *t* within document *d* and *dlen* denotes the length of *d* (sum of all $f_{t'd}$ for all t' in *d*) then we define the local weight

$$lw_{td} = 1/\left(1+exp\left(\alpha \cdot dlen\right)\cdot \lambda^{f_{td}-1}\right) \qquad (1)$$

where $\alpha = 0.0044$ and $\lambda = 0.7$. This formula is derived from the Poisson model of term frequencies within documents (unpublished) and has been found to give good performance on MEDLINE documents.

Baseline results. We here give results of applying the UBW, OBW, and IDF (inverse document frequency[20]) global weights in combination with the local weights (1) as compared with no local weights. Only words from title and abstract indexing are considered in all three test sets.

Table 1. Results of retrieval without (-) and with (+) local weighting. Results are the standard 11-AvgP except for Test Set B where precision over the top 20 ranks (P_{20}) is used because of the nature of the test set[16].

Database	lw_{dt}	IDF	UBW	OBW
OHSUMED (11-AvgP)	-	0.152	0.186	0.214
	+	0.187	0.215	0.225
2,344 MED (11-AvgP)	-	0.491	0.468	0.549
	+	0.520	0.516	0.565
Test Set B (P_{20})	-	0.511	0.540	0.567
	+	0.516	0.547	0.572

First, these results show improvement from local weighting in all cases and this justifies their use in this study. Second, UBW weighting is superior to IDF weighting on two of the test sets. On the third (2,344 MED) UBW is less than 1% below IDF weighting. Because UBW weighting is based on a well founded statistical theory and because it is in all cases near ideal we prefer it for testing indexing quality. This will make our results more model independent. Of course OBW gives better results, but it is over trained. It serves here only as an upper bound on the true ideal that UBW approaches.

AUTOMATIC MESH INDEXING

The Indexing Initiative System (IIS). The NLM Indexing Initiative is a research effort undertaken to explore indexing methodologies for both semi-automated, user-assisted indexing and also for fully automatic indexing applications such as the one described in this paper. The project has created a system, IIS[21], for producing recommended indexing terms for arbitrary biomedical text, especially titles and abstracts of journal articles. The system consists of software for applying alternative methods of discovering MeSH headings and then combining them into an ordered list of recommended indexing terms as shown in Figure 1.

The top portion of the diagram consists of three paths, or methods, for creating a list of recommended indexing terms: MetaMap Indexing, Trigram Phrase Matching, and PubMed Related Citations. The two left paths actually compute UMLS Metathesaurus® concepts which are passed to the Restrict to MeSH method. The results from each path are weighted and combined using the Clustering method. The system is highly parameterized not only by path weights but

also by several internal parameters specific to the Restrict to MeSH and Clustering methods. A brief description of each component follows.

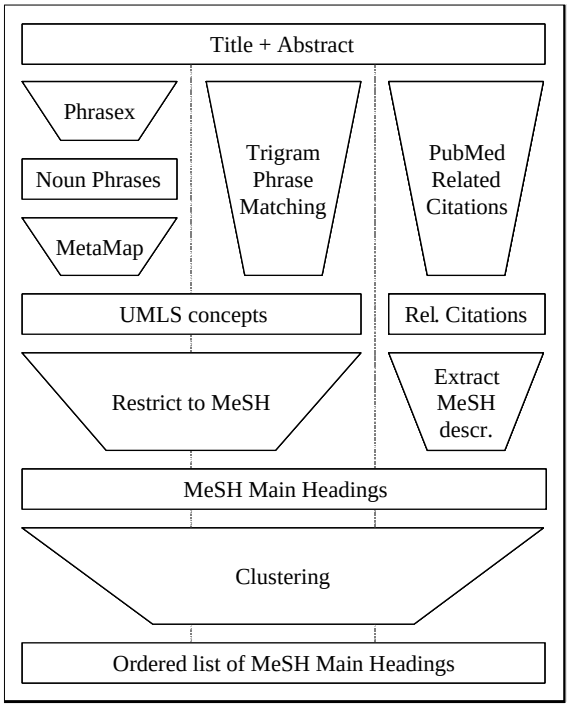

Figure 1. The Indexing Initiative System

The **MetaMap Indexing** (MMI) method of discovering UMLS concepts consists of applying the MetaMap program[22] to a body of text and then ordering the resulting concepts using a ranking function. MetaMap finds Metathesaurus concepts by parsing the text into simple noun phrases using the SPECIALIST(tm) minimal commitment parser, aggressively generating variants for words in the phrase[23], retrieving the candidate set of all Metathesaurus strings containing at least one of the variants, evaluating each candidate against the text using a linguistically principled evaluation metric, and finally constructing a complete mapping for cases when a single concept does not exhaust the text. A list of indexing recommendations is produced from the concepts found in the text using a ranking function emphasizing frequency of occurrence, presence in the title, and MeSH tree depth.

Trigram Phrase Matching is a method of identifying phrases that have a high probability of being synonyms. It is based on representing each phrase by a set of character trigrams that are extracted from that phrase. The character trigrams are used as key terms in a representation of the phrase much as words are used as key terms to represent a document. The similarity of phrases is then computed using the vector cosine similarity measure. Like MMI, the Trigram Phrase Matching algorithm produces UMLS concepts which are subsequently restricted to MeSH headings by the Restrict to MeSH method.

The **Restrict to MeSH** method is based on the observation that the representation of meaning in the UMLS is organized according to the principle of semantic locality[24] in which several means of representing relationships between concepts conspire to produce a cluster of semantically-related terms. In the Indexing Initiative, three of these phenomena are used to find the MeSH terms most closely related to any given UMLS concept: synonyms, interconcept relationships, and categorization[25].

The **PubMed Related Citations** method directly computes a ranked list of MeSH headings based on a given title and abstract. The neighbors of a pending document (related citations) are those documents in the database that are the most similar to it. The similarity between documents is measured by the words (in title and abstract) they have in common using IDF global term weights and local term weights as in (1). This is an example of vector inner product scoring in the paradigm originated by Gerard Salton[20]. Our approach differs from other approaches in how we calculate the local weights. After a pending document has been used to score all database documents, the ranked list is used in a K-Nearest Neighbors method[1, 2] to rank the MeSH terms that are candidates for indexing the document.

Finally, the **Clustering** algorithm produces a single list of recommended MeSH terms by combining the recommendations of the methods described above. It computes a rank score for each suggested indexing term using term weights, co-occurrence information, and estimates of the importance of the term based on where and how the term arose. The result of the clustering process constitutes the output of the IIS.

RESULTS

We applied the UBW with local weighting according to formula (1) to the three test sets with indexing produced by several different schemes. First all the text in titles and abstracts was used (*t*). We also performed indexing based on all the text in humanly assigned MeSH terms (*m*) and again with all the text terms produced by subsets of the automatically assigned MeSH terms (*a*). Finally, we combined text and standard MeSH (*t*+*m*) and text and automatic MeSH (*t*+*a*). The results are given in Table 2. Automatic MeSH consists of the top 5 terms for OHSUMED and 2,344 MED and the top 25 terms for Test Set B. The numbers of terms used here were optimal, presumably because queries are very short in

OHSUMED and 2,344 MED and quite long in Test Set B.

From these results we see that *a* alone is not as good as *m* alone on OHSUMED and 2,344 MED, but slightly better than *m* alone on Test Set B. However *a* appears almost equal to *m* in augmenting *t* on OHSUMED and better than *m* in augmenting *t* on 2,344 MED and Test Set B.

Table 2. Performance of different indexing strategies applied to the three test sets.

Indexing Strategies	OHSUMED 11-AvgP	2,344 MED 11-AvgP	Test SetB P_{20}
t	0.215	0.516	0.547
m	0.150	0.393	0.516
$t+m$	0.245	0.527	0.554
a	0.119	0.370	0.536
$t+a$	0.239	0.544	0.555

DISCUSSION

The research we report here involves two aspects. First, a new method of testing index terms for quality, and second, the results of this testing as applied to the output of the Indexing Initiative System at the NLM. Both aspects deserve comment.

Indexing is generally tested using some retrieval algorithm and almost all the retrieval algorithms that have proved generally useful are based on weighting index terms independently of each other. Such algorithms seek to produce a separation of the documents in the database based on the weights of the terms they contain. The Naïve Bayesian machine learning algorithm also has the purpose to produce a separation of a database into two classes and it can be turned to the same purpose as a retrieval algorithm. In addition in the test set environment the Naïve Bayesian algorithm can produce near ideal results under the independence assumption. This is what leads us to use it for retrieval testing of index terms as described here. The only problem that arises is the overtraining problem and the UBW model solves this.

The true ideal weighting of terms in the Naïve Bayesian paradigm is obtained by using exact probabilities of occurrence of a term within the class of relevant documents and the class of nonrelevant documents of a training set. If the weights are produced from a training set and then applied to the same training set we will generally find them more effective than if they are applied to some new set of documents. This is the overtrained or OBW model. We can avoid overtraining if we remove a single target document from the training set and compute the weights of all the target's terms from the remaining data and then score the target. In this way the target document is scored without using knowledge of its classification and overtraining is avoided. In particular one can see that if the target document is relevant the weights of all its terms will be decreased (score decreased) by removing it and if it is nonrelevant then the weights of all its terms will be increased (score increased) by removing it. We have systematically undertrained the weights specifically for that document. This is the undertrained or UBW model. The UBW model will always give a performance below OBW and the true ideal retrieval will lie between the two. As the size of the training set becomes large both overtraining and undertraining tend to disappear and UBW approaches OBW in performance. This relationship between UBW and OBW is illustrated in Table 1.

Also in Table 1 it may be observed that while IDF is not as good as UBW on OHSUMED and Test Set B, it slightly outperforms UBW on 2,344 MED. While this was not expected, we believe it is a consequence of two factors. First, UBW is undertrained and the database is small so this undertraining is significant. Second, because the database is small it is a less demanding task to separate the relevant from the non-relevant documents in 2,344 MED and a method such as IDF may already be close to adequate (IDF is related to probabilistic models[26]).

The second aspect of the work reported here is the attempt to answer the question whether automatically assigned MeSH terms can provide value in a retrieval environment as defined by several of the test sets of MEDLINE documents that are available for study. Here our principal finding is that the automatically assigned MeSH from the IIS project compare very favorably with the standard humanly assigned MeSH descriptors that are a part of MEDLINE. The indexing *t+a* is only 2% below *t+m* on OHSUMED. On 2,344 MED *t+a* is 3% better than *t+m* while on Test Set B *t+a* and *t+m* are essentially equal. None of the differences are statistically significant. It is true that *m* is significantly better than *a* alone on OHSUMED but this difference is made up by the text (*t*) itself.

While these results are gratifying they only measure one aspect of MeSH indexing and must be viewed with some caution. First, they may be influenced by the fact that the human judges employed in making the test set judgments in most cases have only examined the MEDLINE record in making their judgments. This may have a tendency to devalue the humanly assigned MeSH terms that are based on an examination of the full text of documents. Second, our results do not prove that a human searcher using MeSH terms for Boolean queries as intended would find the automatic MeSH as useful as the humanly assigned MeSH. These concerns could be addressed, at least partially, by incorporating full text into the indexing and test set construction processes on the one hand and by performing user testing in a real-world environment on the other.

REFERENCES

1. Mitchell TM. Machine Learning. Boston: WCB/McGraw-Hill, 1997.
2. Yang Y, Liu X. A re-evaluation of text categorization methods. 22 Annual ACM Conference on Research and Development in Information Retrieval. Berkeley, CA: ACM Press, 1999:42-49.
3. Dumais S, Platt J, Heckerman D, Sahami M. Inductive learning algorithms and representations for text categorization. In: Gardarin G, French J, Pissinou N, Makki K, Bouganim L, eds. Proceedings of the 1998 ACM CIKM International Conference on Information and Knowledge Management. Bethesda, MD: ACM Press, 1998:148-155.
4. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Machine Learning: Proceedings of the Fourteenth International Conference (ICML'97), 1997:412-420.
5. Wilbur WJ. Boosting Naive Bayesian Learning on a Large Subset of MEDLINE. American Medical Informatics 2000 Annual Symposium. Los Angeles, CA: American Medical Informatics Association, 2000:918-922.
6. Mladenic D. Feature subset selection in textlearning. 10th European Conference on Machine Learning (ECML98), 1998:95-100.
7. Kim WG, Wilbur WJ. Corpus-based statistical screening for content-bearing terms. Journal of the American Society for Information Science 2001;52(3):247-259.
8. Fagan JL. The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. Journal of the American Society for Information Science 1989;40(2):115-132.
9. Salton G. Developments in automatic text retrieval. Science 1991;253:974-980.
10. Lewis DD, Sparck Jones K. Natural language processing for information retrieval. Communications of the ACM 1996;39(1):92-101.
11. Hersh W, Buckley C, Leone TJ. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Croft WB, van Rijsbergen CJ, eds. Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland: Springer-Verlag, 1994:192-201.
12. Srinivasan P. Optimal document indexing vocabulary for MEDLINE. Information Processing & Management 1996;32(5):503-514.
13. Haynes RB, McKibbon KA, Walker CJ, Ryan N, Fitzgerald D, Ramsden MF. Online acess to MEDLINE in clinical settings. Annals of Internal Medicine 1990;112:78-84.
14. Hersh WR, Hickam DH, Haynes RB, McKibbon KA. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. Journal of the American Medical Informatics Association 1994;1(1):51-60.
15. Wilbur WJ. The knowledge in multiple human relevance judgments. ACM Transactions on Information Systems 1998;16(2):101-126.
16. Wilbur WJ. A comparison of group and individual performance among subject experts and untrained workers at the document retrieval task. Journal of the American Society for Information Science 1998;49(6):517-529.
17. Langley P, Iba W, Thompson K. An analysis of Bayesian classifiers. Tenth National Conference on Artificial Intelligence. San Jose: AAAI Press, 1992:223-228.
18. Langley P, Sage S. Induction of selective Bayesian classifiers. Tenth Conference on Uncertainty in artificial intelligence. Seattle, WA: Morgan Kaufmann, 1994:399-406.
19. Langley P. Elements of Machine Learning. San Francisco: Morgan Kaufmann Publishers, Inc., 1996.
20. Salton G. Automatic Text Processing. Reading, Massachusetts: Addison-Wesley Publishing Company, 1989. Addison-Wesley Series in Computer Science;
21. Aronson AR, Bodenreider O, Chang HF, et al. The NLM indexing initiative. American Medical Informatics 2000 Annual Symposium. Los Angeles, CA: American Medical Informatics Association, 2000:17-21.
22. Aronson AR, Rindflesch TC, Browne AC. Exploiting a large thesaurus for information retrieval. RIAO 94. Rockefeller University, New York, N. Y: JOUVE, Paris, 1994:197-216.
23. Aronson AR. The effect of textual variation on concept-based information retrieval. In: Cimino JJ, ed. AMIA Annual Fall Symposium. Washington, D. C.: Hanley & Belfus, Inc., 1996:373-377.
24. McCray AT, Nelson SJ. The representation of meaning in the UMLS. Methods of Information in Medicine 1995;34(1-2):193-201.
25. Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In: Lorenzi NM, ed. AMIA Annual Fall Symposium, 1998. Washington, D.C.: Hanley & Belfus, Inc., 1998:815-819.
26. Croft WB, Harper DJ. Using probabiliistic models of document retrieval without relevance information. Journal of Documentation 1979;35(4):285-295.