A Strategy for Assigning New Concepts in the MEDLINE Database

Won Kim[†], PhD and W. John Wilbur[†], MD, PhD

**[†]National Center for Biotechnology Information (NCBI)
National Library of Medicine, Bethesda, MD 20894**

The MeSH® indexing done in MEDLINE® is engineered by humans. Humans define the MeSH concepts and human indexers assign MeSH terms to MEDLINE records. Methods have been designed in an attempt to assign MeSH terms to MEDLINE documents automatically with some success. Methods have also been designed to locate useful phrases as potential concepts for indexing. However, little work has been done on the problem of how one might automatically index with the concepts represented by such phrases. Here we examine this issue and present a method for such indexing.

INTRODUCTION

A good deal of work has been done in an attempt to automatically assign MeSH terms to documents[1-7]. The methods developed generally take advantage of having already a set of documents where indexing has been done and learning from this prior indexing. Work has also been directed towards automatically identifying good phrases that may represent important concepts for indexing in biomedicine[8-10]. Given that one has collected a group of phrases that might serve as the basis of concepts by which to index literature, there is a second problem, identifying phrases which are synonyms that require grouping under the same concept. In this area also there has been useful work[5, 8, 11]. Our interest here is in a third, but equally important, step. This step deals with automatically assigning the concepts, that would result from the first two steps, to documents. This is particularly challenging as this scenario cannot make use of a repository of training documents where the concepts have already been humanly applied (that would not be a truly automatic method). We examine several methods and show that the best of these falls only a little behind a method of training on data with previously assigned MeSH terms.

METHODOLOGY

Our approach is to examine ten randomly chosen MeSH terms and the phrases representing their underlying concepts. For each MeSH term we attempt to reconstruct the MeSH assignments of that term in MEDLINE making use of nothing more than the group of synonymous phrases which represent the underlying concept in the UMLS® Metathesaurus. The ten MeSH terms are chosen from the leaf nodes in the MeSH trees in order to avoid some of the complexity that arises from higher level nodes. Higher level nodes will be considered subsequently in our discussion. Among the leaf nodes there are a few cases where the MeSH concept seems to have little correlation with the occurrence of the phrases representing the concept. We have also chosen to avoid such cases. We do this by insisting on an overlap or Dice coefficient of at least 0.5 between the MeSH assignments and the documents that contain one of the concept phrases. If x is a MeSH term let M_x stand for the documents in all of MEDLINE where

Table 1. Ten MeSH terms chosen for study and the Dice coefficients between the MeSH and their corresponding English concept phrases.

MeSH Term x	$\lvert M_{cx} \rvert$	$\lvert M_x \rvert$	$\lvert M_{cx} \cap M_x \rvert$	Dice Coef
Anorexia Nervosa	5847	7082	4818	.75
Autistic Disorder	7129	7144	5693	.80
Blepharophimosis	306	142	116	.52
Cardiomyopthy Congestive	8008	7824	5038	.65
Dermatitis Herpetiformis	1467	1728	1059	.66
Maleic Hydrazide	190	94	85	.60
Rinderpest	700	452	348	.60
Scrapie	2594	1955	1568	.69
Substantia Nigra	11030	9455	5793	.57
Thiosulfate Sulfurtranferase	580	496	407	.76

this term is assigned and M_{cx} stand for the documents in all of MEDLINE where one of the phrases representing x appears in either the title or the abstract. Then the Dice coefficient is

$$2\left|M_x \cap M_{cx}\right| / \left(\left|M_x\right| + \left|M_{cx}\right|\right) \qquad (1)$$

where $\left|X\right|$ denotes the number of elements in the set X. The MeSH terms we chose are listed in Table 1.

Boolean Prediction. Here the set M_{cx} represents our prediction. This set is the result set of a Boolean 'OR' of all the phrases representing the concept cx corresponding to the MeSH term x. For purposes of evaluation we assign a score of 1 to each document in M_{cx} and 0 to all other documents.

Vector Retrieval Prediction. We are dealing with natural language text and we prepare it all in the same manner. Stop words are removed, but no stemming is done. In addition to single words, we also include two word phrases without punctuation or stop words. No MeSH terms are included. In this way each MEDLINE document is given a bag-of-words representation based on its title and abstract. The same preparation is applied to the natural language phrases representing a MeSH term x to produce a bag-of-words document which we will represent by q_x. We then apply TF×IDF weighting to do vector retrieval with each q_x as query and against all of MEDLINE. 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 or TF factor by

$$tf_{td} = 1 / \left(1 + exp\left(\alpha \cdot dlen\right) \cdot \lambda^{f_{td} - 1}\right) \qquad (2)$$

where $\alpha = 0.0044$ and $\lambda = 0.7$ [12]. The global weight or IDF factor is given by the relatively standard[13, 14]

$$IDF_t = \log\left(N / n_t\right) \qquad (3)$$

where n_t is the number of documents in MEDLINE containing the term t and N represents the size of MEDLINE ($\cong$ 15 million). The scores resulting from a query q_x rank all the documents in MEDLINE according to their likelihood of having the MeSH term x assigned.

Naïve Bayesian Prediction. Here we use naïve Bayesian machine learning in an attempt to predict when a MeSH term is assigned. However our approach will be nonstandard because we do not allow ourselves a training set where the MeSH term has already been assigned. Rather we have the set M_{cx} from which we can try to learn how the MeSH term is assigned. We choose the Baysian approach for two reasons. First, because it is efficient to use on the very large MEDLINE collection where few other methods can even be applied and none is efficient. Second, because naïve Bayes is robust under errors in the training set. A small fraction of the training set can be mislabeled and the Bayesian weights will change but little, while if a few support vectors for a Support Vector Machine are mislabeled it can have a profound effect.

With the use of the Bayesian method there comes the opportunity to do feature selection. Work by Yang and Pedersen[15] suggests that up to 90% of features are not necessary. Other research suggests that a threshold on the Bayesian weights may be the most effective way to prune away useless features[16, 17]

We have implemented four different methods of feature selection. We apply them to the single word and two word phrase features described under Vector Retrieval Prediction. For any set of documents X let $\bar{X}$ denote the complement of X in MEDLINE. For an arbitrary term t let M_t denote the set of documents in MEDLINE that contain the term t in their title or abstract. Then we need certain document counts to define feature selection measures. Define

$$n_x = \left|M_{cx}\right|,\ n_t = \left|M_t\right|,\ n_{\hat{x}} = \left|\bar{M}_{cx}\right|,\ n_{\hat{t}} = \left|\bar{M}_t\right| \qquad (4)$$

and in a similar manner

$$\begin{aligned} n_{tx} &= \left|M_t \cap M_x\right|,\ n_{t\hat{x}} = \left|M_t \cap \bar{M}_x\right|, \\ n_{\hat{t}x} &= \left|\bar{M}_t \cap M_x\right|,\ n_{\hat{t}\hat{x}} = \left|\bar{M}_t \cap \bar{M}_x\right|. \end{aligned} \qquad (5)$$

Then the different measures are given by

1) Bayesian weight

$$BW_t = \log\left(\frac{n_{tx} n_{\hat{t}\hat{x}}}{n_{t\hat{x}} n_{\hat{t}x}}\right) \qquad (6)$$

2) Log of Chi Square

$$L\chi_t^2 = \log\left(\frac{N\left(n_t n_x - n_{tx} N\right)^2}{n_t n_x n_{\hat{t}} n_{\hat{x}}}\right) \qquad (7)$$

3) Mutual Information

$$\begin{aligned} MI_t = n_{tx} \log\left(\frac{n_{tx} N}{n_t n_x}\right) + n_{t\hat{c}_x} \log\left(\frac{n_{t\hat{x}} N}{n_t n_{\hat{x}}}\right) \\ + n_{\hat{t}x} \log\left(\frac{n_{\hat{t}x} N}{n_{\hat{t}} n_x}\right) + n_{\hat{t}\hat{c}_x} \log\left(\frac{n_{\hat{t}\hat{x}} N}{n_{\hat{t}} n_{\hat{x}}}\right) \end{aligned} \qquad (8)$$

4) Dice Coefficient

$$DC_t = \frac{2 n_{tx}}{n_t + n_x} \qquad (9)$$

In this study, Bayesian learning with these four feature selection strategies is applied to learn the difference between the sets M_{cx} and $\bar{M}_{cx}$ with the aim to predict the members of M_x. Learning the concept here is manifested by an estimate for the optimal threshold for a particular feature selection strategy as well as the Bayesian term weights for the terms that satisfy that threshold.

Evaluation. Because we use methods that rank all the documents in MEDLINE and attempt to rank documents from M_x above documents from $\bar{M}_x$, it is convenient to score the results as an *average precision at seen relevant documents*. An *average precision at seen relevant document*s is the average of precisions obtained at the points where each relevant document (member of M_x) is observed in the ranking.

RESULTS

Baseline results. We here give the results of applying Boolean and vector retrieval based on the concept phrases representing the concept cx corresponding to a MeSH term x.

Table 2 Baseline results for Boolean and Vector Retrieval.

	Average Precision	
MeSH Term x	Boolean cx	Vector Retrieval
Anorexia Nervosa	.562	.665
Autistic Disorder	.637	.501
Blepharophimosis	.319	.460
Cardiomyopthy Congestive	.406	.525
Dermatitis Herpetiformis	.443	.500
Maleic Hydrazide	.417	.305
Rinderpest	.386	.297
Scrapie	.486	.630
Substantia Nigra	.323	.501
Thiosulfate Sulfur-transferase	.578	.353
Average	.456	.474

Vector retrieval proves superior to the Boolean 'OR' over the phrases representing the concept in six of ten cases and the overall average favors vector retrieval.

Bayesian learning with feature selection. MeSH terms are assigned by human experts and we hope to approximate how a human expert may assign the MeSH terms to MEDLINE documents. We argue that when a human expert is given the task whether a MeSH terms should be assigned to a MEDLINE document, he or she makes the judgment based upon relatively few features or terms that are relevant to the MeSH concept in the given document. We considered four different feature selection strategies to extract the salient features relevant to the concept. For each selection strategy and each MeSH term, we apply naïve Bayes to learn the full set of weights and then the selection strategy to determine the optimal threshold for that MeSH term using that strategy. We then average the optimal thresholds from the other nine MeSH terms for the given strategy and apply the result to the term under investigation. This form of cross validation allows us to make predictions for the assignment of a MeSH term x that do not depend in any way on knowledge of M_x. Results for the four different strategies of feature selection are given in Table 3.

Table 3 Bayesian learning with the four different feature selection strategies.

	Average Precision			
MeSH Term x	MI_t	DC_t	$L\chi_t^2$	BW_t
Anorexia Nervosa	.706	.666	.670	.668
Autistic Disorder	.734	.752	.742	.746
Blepharophimosis	.630	.463	.585	.615
Cardiomyopathy Congestive	.485	.610	.601	.606
Dermatitis Herpetiformis	.521	.480	.501	.499
Maleic Hydrazide	.385	.414	.414	.575
Rinderpest	.356	.356	.382	.400
Scrapie	.378	.355	.381	.460
Substantia_Nigra	.505	.514	.543	.518
Thiosulfate Sulfurtransferase	.664	.701	.708	.727
Average	.537	.531	.553	.581

We see that the results are for all four methods markedly better than the results of the baseline calculations given in Table 2. Further, feature selection based on Bayesian weights gives the best overall performance. Table 4 gives the average number of terms used by the different feature selection methods in learning a concept. It is evident that there are substantial differences between the methods in the number of terms they select. Table 5 gives the average threshold used by each of the methods where the average is taken over all ten concepts. The results in Table 3 are based on cross validation where the threshold used for a particular concept is obtained as an average of the optimal thresholds for the other nine concepts. Based on that data we can expect a similar performance from any of the feature selection methods if we use the corresponding threshold in Table 5 for some new concept not involved in this study. Of course that will need to be a concept that has a reasonable overlap (a Dice coefficient of at least 0.5) with the Boolean query result based on the phrases which represent that concept.

Table 4 Average number of features selected by each of the four different strategies in learning to recognize a concept.

	MI_t	DC_t	$L\chi_t^2$	BW_t
Total terms	29.7	8.7	37.0	1046.6
Single Title Terms	3.4	1.7	2.7	14
Single Abs Terms	16.6	2.7	7.2	54.2
Phrase Title Terms	1.9	1.1	5.6	146.1
Phrase Abs Terms	7.8	3.2	21.5	832.3

Table 5 Average threshold used by the four different strategies in learning to recognize a concept.

MI_t	DC_t	$L\chi_t^2$	BW_t
23349	0.268	6.887	9.64

DISCUSSION

First, it is important to note that, even though our data set is small, we have significant results from our study. A comparison of the results in Table 2 and Table 3 show that the naïve Bayesian learning with feature selection based on a threshold for Bayesian weights is superior to the results obtained by vector retrieval based on the phrases belonging to the UMLS concept in nine out of ten cases. By the sign test[18] this result is significant with a p-value of 0.02. From this we conclude that there is definite value in the naïve Bayesian approach with feature selection to predict the assignment of a concept to MEDLINE records. Based on the limited data reported here there is not sufficient evidence to conclude that one of these feature selection methods is superior to the others. In order to elucidate this issue we selected an additional twenty MeSH terms and corresponding concepts and performed the same analysis that is reported here. We found among the aggregate total of thirty cases that the Bayesian weight approach was superior to the mutual information approach in 26 of 30 cases, to the Dice coefficient method in 20 of 30 cases, and superior to the Chi square method in 23 of 30 cases. The sign test applied to each of these cases shows the Bayesian weight method to be superior to the other method at the five percent significance level. In all of the additional twenty cases we used the thresholds determined to be optimal from the first ten cases and which are reported in Table 5. This significantly reduced the amount of calculation necessary for these additional tests. Due to space limitations we are unable to show the details of the additional twenty cases.

An important question is how well are we doing in predicting the MeSH assignments. One guage is how well we can predict the MeSH assignments if we use the same naïve Bayesian learning method with feature selection based on BW_t and apply it to learn to distinguish M_x and $\bar{M}_x$, i.e., the standard machine learning approach. We did this for the ten MeSH terms studied here and found an overall average precision of 0.650. This is 12% better than the results of learning on M_{cx} and $\bar{M}_{cx}$ (last column of Table 3). Part of this difference is due to the fact that different concepts have different optimal thresholds for selecting features. Thus when we learn from nine and extrapolate to the tenth term we are generally not able to give an optimal threshold. In order to illustrate this point we averaged the nine cases and used the result as the threshold for the tenth case for standard machine learning and our overall average precision dropped from 0.650 to 0.627 or a drop of 4%. Thus we could likely improve our performance if we could find a better way to predict the optimal threshold for a given concept.

Finally, there is the question of how broadly we may apply the method described here. Not all MeSH terms are leaf nodes in the MeSH trees. However, nodes that are not leaf nodes may be considered the sum of all their leaves and if we can make useful assignments for the leaves this defines the assignments of the higher level nodes. But here we have not actually shown how to deal with all the leaves. We have required a leaf node as a MeSH concept to have a Dice coefficient with the set of documents that contain one of the phrases defining the concept of at least 0.5. Roughly we are saying that if about half the documents in the training set would have the MeSH term assigned and about half of the documents that would have the MeSH term assigned are in the training set, then we can learn at a reasonable level how to assign the MeSH term based on that set of documents. In this study we have used the phrases belonging to the MeSH concept to define such a set. But for some MeSH concepts this strategy will not give such a set. And in general for a new concept not already a part of UMLS one obviously must resort to other means to find the initial set of documents which are a rough approximation to what one wishes to learn.

CONCLUSION

We have presented a method to predict the assignment of a new concept to MEDLINE documents. The method is based on naïve Bayesian learning on a set that is a rough approximation of the target assignments which one seeks. Thus it is a form of bootstrapping. We show that it is superior to a baseline vector retrieval method and only about 12% less accurate than the standard machine learning approach with naïve Bayes' which one can apply when one has a training set where the concept is already assigned. In future work we plan to refine the method and find effective methods of assembling synonymous phrases to use in approximating new concepts that might be candidates for new MeSH terms.

References

1. 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.
2. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo 2004;2004:268-72.
3. Cooper GF, Miller RA. An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text. J Am Med Inform Assoc 1998;5(1):62-75.
4. Fowler J, Maram S, Kouramajian V, Devadhar V. Automated MeSH indexing of the World-Wide Web. Proc Annu Symp Comput Appl Med Care 1995:893-7.
5. Hahn U, Honeck M, Piotrowski M, Schulz S. Subword segmentation--leveling out morphological variations for medical document retrieval. Proc AMIA Symp 2001:229-33.
6. Kouramajian V, Devadhar V, Fowler J, Maram S. Categorization by reference: a novel approach to MeSH term assignment. Proc Annu Symp Comput Appl Med Care 1995:878-82.
7. Yang Y. An evaluation of statistical approaches to MEDLINE indexing. Proc AMIA Annu Fall Symp 1996:358-62.
8. Nenadic G, Spasic I, Ananiadou S. Terminology-driven mining of biomedical literature. Bioinformatics 2003;19(8):938-43.
9. Kim WG, Wilbur WJ. Corpus based statistical screening for phrase identification. Journal of the American Medical Informatics Association 2000;7:499-511.
10. 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.
11. Wolff-Terroine M, Rimbert D, Rouault B. Improved statistical methods for automatic construction of a medical thesaurus. Methods Inf Med 1972;11(2):104-13.
12. Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. Proc. AMIA Symp. Washington, D.C., 2001:319-324.
13. Salton G. Automatic Text Processing. Reading, Massachusetts: Addison-Wesley Publishing Company, 1989. Addison-Wesley Series in Computer Science;
14. Witten IH, Moffat A, Bell TC. Managing Gigabytes. (Second ed.) San Francisco: Morgan-Kaufmann Publishers, Inc., 1999.
15. 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.
16. Mladenic D. Feature subset selection in textlearning. 10th European Conference on Machine Learning (ECML98), 1998:95-100.
17. Mladenic D, Grobelnik M. Feature selection for unbalanced class distribution and naive Bayes. Sixteenth International Conference on Machine Learning: Morgan Kaufmann Publishers Inc., 1999:258-267.
18. Larson HJ. Introduction to Probability Theory and Statistical Inference. (3 rd ed.) New York: John Wiley & Sons, 1982.