# A semi-supervised learning method to classify grant support zone in Web-based medical articles

Xiaoli Zhang*, Jie Zou, Daniel X. Le, George Thoma
Communications Engineering Branch
National Library of Medicine, NIH, Bethesda MD 20814, USA
{zhangxiaol, jzou, danle, gthoma}@mail.nih.gov

## ABSTRACT

Traditional classifiers are trained from labeled data only. Labeled samples are often expensive to obtain, while unlabeled data are abundant. Semi-supervised learning can therefore be of great value by using both labeled and unlabeled data for training. We introduce a semi-supervised learning method named *decision-directed approximation* combined with Support Vector Machines to detect zones containing information on grant support (a type of bibliographic data) from online medical journal articles. We analyzed the performance of our model using different sizes of unlabeled samples, and demonstrated that our proposed rules are effective to boost classification accuracy. The experimental results show that the decision-directed approximation method with SVM improves the classification accuracy when a small amount of labeled data is used in conjunction with unlabeled data to train the SVM.

**Keywords:** semi-supervised learning, Support Vector Machines (SVM), decision-directed approximation

## 1. INTRODUCTION

In many modern classification problems such as text categorization and information retrieval, it often happens that few labeled samples are available but a large pool of unlabeled samples can be easily acquired. Various methods have been proposed to improve classification performance by taking advantage of unlabeled data. These include: Expectation-Maximization (EM) with generative mixture models, self training, co-training, transductive support vector machines, and graph-based methods, etc. Nigam et al. applied EM combined with a naïve Bayesian classifier to estimate maximum a posteriori parameters from labeled and unlabeled text documents[1]. Yarowsky used self-training for word sense disambiguation in a given context[2]. Blum et al. used co-training on Web-page classification[3]. In this method, two classifiers are trained with the labeled data on two disjoint sub-feature sets separately, and then each learns from the other classifier's predictions on the new unlabeled samples. Joachime used transductive inference for SVM by taking into consideration the test data for learning margins[4].

Locating zones in an article is a preliminary step before extracting bibliographic data. In this paper we introduce a semi-supervised learning method – *decision-directed approximation* (self training) with SVM to identify zones containing a particular kind of bibliographic item (grant support, defined in Section 2) from Web-based medical journal articles. The challenge with grant support zone detection is that conventional classifiers are limited due to the insufficiency of labeled training samples. Of all the types of grant supports, some associate with well-formatted grant numbers and zones containing grant numbers are easier to extract through string matching[5]. This renders the possibility of using available labeled grant number zones plus unlabeled zones for grant support zone classification. Our semi-supervised SVM method shows higher accuracy compared to an SVM classifier trained only from grant number zones.

The rest of this paper is organized as follows. In Section 2, we explain the grant support information encountered in the MEDLINE database, and present the problem. We describe SVM classifier for zone labeling and decision-directed approximation in Sections 3 and 4 respectively. In Section 5 we analyze the performance of the proposed method with experimental results. We conclude and summarize our work in Section 6.

*Xiaoli Zhang: zhangxiaol@mail.nih.gov, phone 1 -301- 435-3245.

## 2. GRANT SUPPORT

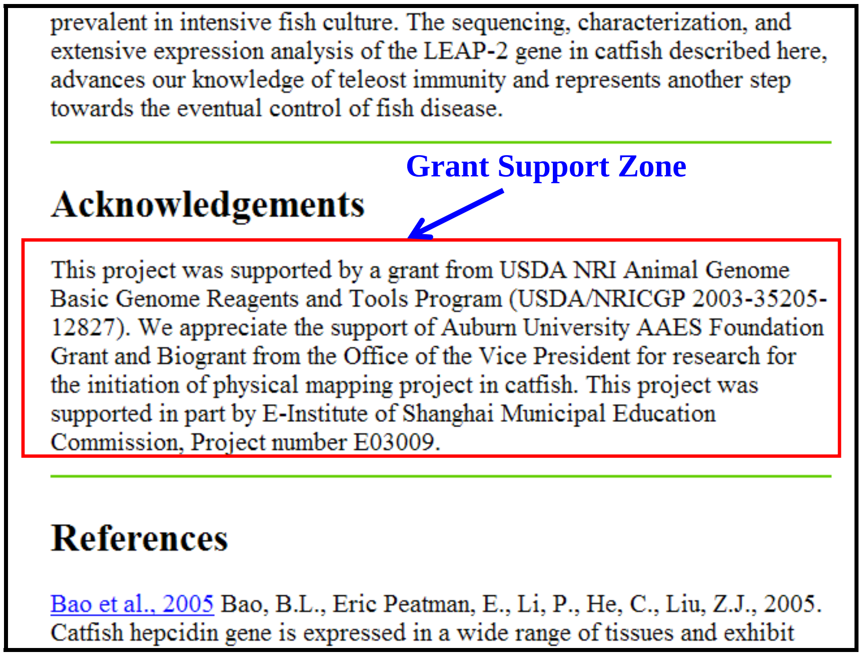
MEDLINE ®, the flagship database of the U.S. National Library of Medicine, contains 17 million citations to the medical journal literature. Given the ever-increasing volume of medical journal articles published in HTML and PDF formats and high labor cost of manual entry, automatic extraction of bibliographic data such as title, author, affiliation, grant support is crucial for building citations for MEDLINE. In this work, we focus on identifying segmented zones in online articles that contain information on grant support.

*Grant Support (GS)*

GS is a required field in a MEDLINE citation, referring to the type of organization that supports the research reported in an article. At present, six types of grant supports exist, as defined in Table 1. Grant supports usually appear in a paragraph as organization names with some "support words" such as "supported", "funded", "financed", "grant", and so on[5]. Our task is to identify zones containing grant supports, using as clues the "support words" if they exist in the zone.

Table 1. Six types of grant supports

| **Grant Support Type** | **Definition** |
|---|---|
| Non-U.S. Gov't | Support from universities, companies, private institutions, foreign countries, etc. |
| U.S. Gov't, Non-P.H.S. | Support from US government, other than PHS organizations. |
| U.S. Gov't, P.H.S | Support from one of the PHS organizations such as AHRQ, ATSDR, CDC, DHHS, FDA, HRSA, INS, NIH, OASH, SAMHSA, and VA. |
| NIH Extramural | Support from an institute or center of the National Institutes of Health. |
| NIH Intramural | Support from one of the NIH organizations for intramural research. |
| Wellcome Trust | Support from the Wellcome Trust, a granting institution in the U.K. |



prevalent in intensive fish culture. The sequencing, characterization, and extensive expression analysis of the LEAP-2 gene in catfish described here, advances our knowledge of teleost immunity and represents another step towards the eventual control of fish disease.

Acknowledgements

This project was supported by a grant from USDA NRI Animal Genome Basic Genome Reagents and Tools Program (USDA/NRICGP 2003-35205-12827). We appreciate the support of Auburn University AAES Foundation Grant and Biogrant from the Office of the Vice President for research for the initiation of physical mapping project in catfish. This project was supported in part by E-Institute of Shanghai Municipal Education Commission, Project number E03009.

References

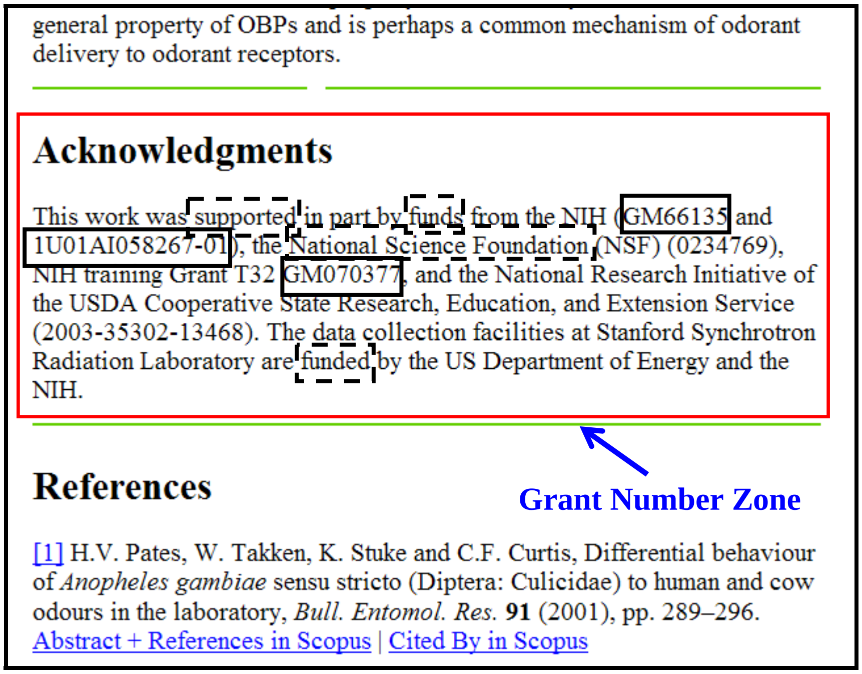
Bao et al., 2005 Bao, B.L., Eric Peatman, E., Li, P., He, C., Liu, Z.J., 2005. Catfish hepcidin gene is expressed in a wide range of tissues and exhibit

Fig. 1. Examples of granting organizations (Non-U.S. Gov't and U.S. Gov't, Non-P.H.S.)

Figure 1 shows examples of funding sources in an article. Here the grants are from USDA (U.S. Gov't, Non-P.H.S.), Auburn University AAES Foundation (Non-U.S. Gov't), and E-Institute of Shanghai Municipal education Commission (Non-U.S. Gov't), respectively.

*Grant Number (GN)*

When the research is funded by one of the institutes at NIH or the U.S. PHS, the article mentions a grant number and the corresponding granting organization (institute). Grant numbers are associated with the funding agencies of the U.S. Gov't P.H.S., NIH Extramural or NIH Intramural grants. A typical GN zone marked with a thick red bounding box is shown in Figure 2. Three grant numbers and some informative words, which are helpful to GN zone detection, are highlighted with solid and dotted boxes respectively. Grant number zone is one type of grant support zone always containing grant numbers.



general property of OBPs and is perhaps a common mechanism of odorant delivery to odorant receptors.

**Acknowledgments**

This work was supported in part by funds from the NIH (GM66135 and 1U01AI058267-01), the National Science Foundation (NSF) (0234769), NIH training Grant T32 GM070377, and the National Research Initiative of the USDA Cooperative State Research, Education, and Extension Service (2003-35302-13468). The data collection facilities at Stanford Synchrotron Radiation Laboratory are funded by the US Department of Energy and the NIH.

**References**

[1] H.V. Pates, W. Takken, K. Stuke and C.F. Curtis, Differential behaviour of *Anopheles gambiae* sensu stricto (Diptera: Culicidae) to human and cow odours in the laboratory, *Bull. Entomol. Res.* **91** (2001), pp. 289–296.
Abstract + References in Scopus | Cited By in Scopus

Fig. 2. An example of GN zone

Labeling grant support zones manually is very labor intensive. However, there are plenty of unlabeled grant support zones available. Due to the well-defined GN formats, GN extraction can be easily implemented through simple string matching. It is difficult to exploit string information for GS support zone detection because of various sources of grant support. Therefore, GN zones are much easier to identify than the other grant support zones. We have some articles with labeled grant number zones and other zones, but many articles without labeled zones. The problem is that we cannot design a classifier to detect zones with different types of grant supports when there are a low number of labeled GS zones, other than the ones that contain grant numbers. The similar "support words" in GN zone and GS zone allow us to train a classifier using a limited subset of grant support zones - GN zones and to adapt the classifier to detect different types of GS zones by taking advantage of the unlabeled zones. A semi-supervised learning approach named decision-directed approximation is thus proposed as a practical solution to bootstrap a classifier by using both labeled and unlabeled data. SVM classifier is adopted in the decision-directed approximation considering the high dimensionality of feature vectors representing article zones. The decision-directed approximation method combined with SVM will be used for grant support zone labeling.

## 3. SVM CLASSIFIER FOR ZONE LABELING

GS zone labeling is preceded by an HTML zoning step, which is to segment the whole HTML article into zones by analyzing the geometric and text features of the zones. HTML zoning is a very useful preprocessing step for several information retrieval tasks, and has been discussed in our previous work [6, 7]. After the HTML article is segmented into zones, we formulate the GS zone labeling as a two-class text categorization problem, i.e., classifying zones into GS zones (the zones containing grant support) and "other" zones (the zones not containing grant support).

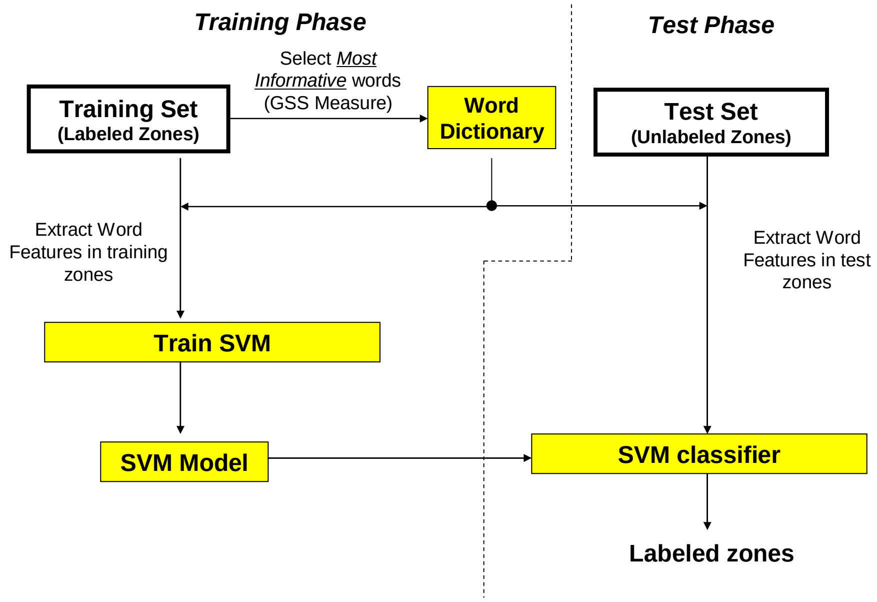


Fig. 3. The training and testing phases of an SVM-based GS zone labeling

We adopt the Support Vector Machine for our GS zone labeling. Figure 3 illustrates the training and testing phases of our SVM-based GS zone labeling method. We use a set of binary features indicating whether or not particular words appear in the zones for the classification. Due to the large amount of words (hundreds of thousands) appearing in the corpus, the first step in the training phase is feature selection, i.e., selecting a set of the most informative words. This is achieved through the GSS measure [8], named after the three authors who proposed the method. In a survey of text categorization by Sebastiani [9], GSS measure is recognized as one of the best methods for feature dimension reduction. In our two-class classification, the GSS measure of a given word $t_k$ is defined as:

$$GSS(t_k) = \left| P(t_k, c_1) P(\bar{t}_k, c_0) - P(t_k, c_0) P(\bar{t}_k, c_1) \right|$$

where $c_0$ and $c_1$ are the labels for GS zones and other zones, respectively, and $P(\bar{t}_k, c_i)$ indicates the probability that, given a random zone, word $t_k$ does not appear in the zone, and that the zone belongs to category $c_i$. The GSS measure reflects the intuition that the best words are the ones distributed most differently in the GS and other zones. $P(t_k, c_i)$ and $P(\bar{t}_k, c_i)$ can be estimated by counting occurrences in the training samples.

Table 2. List of words with the highest GSS measures

| | | | | |
|---|---|---|---|---|
| supported | grant | assistance | institutes | work |
| university | center | acknowledgement | program | award |
| research | thank | foundation | science | health |
| fellowship | discussion | national | manuscript | trust |

All the words can then be sorted according to their GSS measures. A higher value for this measure generally indicates better discriminating ability. Table 2 shows 20 words with the highest GSS measures.

Once the dictionary consisting of the words with the highest GSS measures is selected, a binary feature vector, $\mathbf{f_i} = \{ f(t_1, d_i), \cdots, f(t_k, d_i), \cdots, f(t_n, d_i) \}$, is extracted from each zone. $t_k$ is the k$^{th}$ word in the dictionary, $n$ is the dictionary size, $d_i$ is a zone, and $f(t_k, d_i)$ indicates whether the word $t_k$ appeared in zone $d_i$ or not. These feature vectors serve to represent the zones, and are used to train the SVM classifier. Similarly, word features are extracted from each test zone. The trained SVM then classifies the unlabeled binary feature vectors and predicts the labels of test zones.

## 4. DECISION-DIRECTED APPROXIMATION

Decision-directed approximation is a strategy to update a defined classifier each time an unlabeled sample is classified and then added to the training set [10]. The parameters of the classifier will be re-estimated and in turn the pseudo-labeled sample will be reclassified. Alternatively, this strategy can be applied to update a classifier after all *n* samples are classified. The process can be repeated until the predicted labels stop changing.

Due to the high dimensionality of the feature vectors of article zones, we use SVM classifier in our model. Additionally, we apply several rules to boosting the performance of this model. First, the most confident samples (grant support zones and other zones) with their predicted labels are added to the training set each time after the new unlabeled article zones are classified (*Rule 1*). Second, when updating the training set with the pseudo-labeled article zones, we also update the word dictionary by recalculating the GSS measures of the informative words for grant support (*Rule 2*). Third, instead of taking unlabeled samples all at once, we sequentially add different number of articles with unlabeled zones. At the first time a new set of unknown article zones are encountered, they are classified by SVM trained on the initial ground truth data mixed with previous pseudo-labeled zones selected with high confidence value (*Rule 3*). We start from a small set of training article zones which can be increased by keep adding pseudo-labeled zones. The following steps elaborate the procedure of our algorithm. For brevity, we use samples to represent article zones.

```
1. Input the initial dataset D1 with a small set of training samples;
2. Train SVM classifier;
3. Input a new dataset D2 with unlabeled samples;
4. If the number of all the unlabeled and pseudo-labeled samples is greater than the predefined value, stop;
5. Classify the unlabeled samples in dataset D2 and obtain the pseudo-labels for these samples;
6. Select the most confident pseudo-labeled samples to construct dataset M;
7. Use the new training set D1 + M to update the word dictionary and retrain SVM classifier;
8. Reclassify the pseudo-unlabeled samples in dataset D2;
9. If the number of iterations from 6-8 reaches the maximum count, stop; otherwise repeat steps 6-8 until the pseudo labels won't change many from the previous ones;
10. Update the training set D1 = D1 + M, go to step 3;
```

Fig. 4. Algorithm description of decision-directed approximation

## 5. EXPERIMENTS

### 5.1 Data description

We collected our training and test samples from the MEDLINE 2006 database. There are a total of 660000 medical journal articles, of which 281218 have grant support information, and the rest do not. In Table 3 we show the distribution of five types of grant supports in the articles.

Table 3. Distribution of different GS types in the MEDLINE 2006 database

| Non-U.S. Gov't | U.S. Gov't, Non-P.H.S. | U.S. Gov't, P.H.S. | NIH Extramural | NIH Intramural |
|---|---|---|---|---|
| 68.69% | 7.69% | 1.04% | 21.62% | 0.96% |

Our training set starts from a small set of samples, 100 articles containing grant number and with labeled zones. This initial small sample set makes it convenient for us to observe the tendency of SVM classification by adding unlabeled samples. The dataset of unlabeled samples being added includes 1000 articles each of which was randomly selected from 2006 MEDLINE data with grant support. We also collected another 1000 articles with labeled zones as test data to evaluate the SVM classifier trained by adding unlabeled samples. Out of the 1000 articles, 470 articles have grant supports and they are further separated according to the percentages of five types of grant supports, as shown in Table 3.

### 5.2 Experimental results

We use LibSVM [11], an SVM library developed at National Taiwan University, to implement our GS zone classification. We adopted Radial Basis Function (RBF) as the kernel function where the two parameters, $C$ (penalty parameter of the errors) and $\gamma$ (RBF parameter), are selected through exhaustive grid-search using cross-validation on the training samples [13]. The training step creates a SVM model, which is then used for labeling the test zones.

All the articles involved were first segmented into zones by an HTML journal article segmentation algorithm [6, 7]. For grant support articles, there is averagely one GS zone per article and the rest are other zones (non-GS zones). We collected 100 grant number zones and also randomly selected 100 other zones from the 100 articles with GN as the initial training zones for the SVM classifier. The procedure to add the 1000 articles with unlabeled zones are described as follows. First we classify 10 articles, then add the pseudo-labeled zones into the initial training set, reclassify the 10 articles, and repeat the steps 6-9 in Figure 4. Following the same procedure, we sequentially add 20, 70, 200, and 700 articles. Therefore, the numbers of articles added to the initial training set of 100 articles will be 10, 30, 100, 300 and 1000. Each set of articles are added sequentially, for example, 30 articles are added as 10 + 20, and 100 as 10 + 20 + 70, and so on. The number of unlabeled articles being added increases exponentially. Totally 1000 articles including GS are added to the initial 100-article set at the end. Each time a set of articles are classified, the most confident pseudo-labeled grant support zones and the same number of randomly selected other zones are added into the current set of training zones to update the SVM classifier. To demonstrate that our model with the rules in Section 5 helps GS zone classification, we did the following experiments for comparison:

1. After new articles are classified and pseudo-labeled zones are added to the training set, we do not update the word dictionary, and still use the dictionary constructed from the previous training set to observe the effects of updating the dictionary with newly pseudo-labeled zones (Rule 2*);

2. Instead of adding unlabeled articles sequentially (Rule 3), we apply decision-directed approximation by adding different numbers of unlabeled articles directly to the initial 100 articles with GN;

We evaluated decision-directed approximation with SVM model on 1000 articles, in total 87846 zones with 477 grant support zones and 87369 other zones. The maximum number of iterations to repeat the steps 6-8 in Figure 4 is preset to 50 for obtaining stable classification results. Tables 4-6 showed the performances of our model with different rules and different numbers of articles added. We use precision and recall rates as performance measures for evaluating the classification of 87846 zones.

Table 4. Classification results at different numbers of articles added to the initial 100-article set without using Rule 2

| Number of articles being added | 0 | 10 | 30 | 100 | 300 | 1000 |
|---|---|---|---|---|---|---|
| Precision | 75.12% | 77.05% | 77.98% | 78.86% | 74.92% | 68.92% |
| Recall | 98.11% | 98.11% | 98.32% | 98.53% | 98.32% | 98.74% |

*Rule 1 defined in Section 4 is a general rule for decision-directed approximation method. We discuss only Rule 2 and Rule 3 specifically in our model.

Table 5. Classification results at different numbers of articles added to the initial 100-article set without using Rule 3

| Number of articles being added | 0 | 10 | 30 | 100 | 300 | 1000 |
|---|---|---|---|---|---|---|
| Precision | 75.12% | 80.24% | 79.49% | 74.92% | 74.52% | 69.48% |
| Recall | 98.11% | 98.11% | 98.11% | 97.90% | 97.90% | 97.48% |

Table 6. Classification results at different numbers of articles added to the initial 100-article set with Rules 2 & 3

| Number of articles being added | 0 | 10 | 30 | 100 | 300 | 1000 |
|---|---|---|---|---|---|---|
| Precision | 75.12% | 80.24% | 83.07% | 81.24% | 77.05% | 73.29% |
| Recall | 98.11% | 98.11% | 98.32% | 98.53% | 98.95% | 99.37% |

We can observe from Tables 4-6 that, given the same number of unlabeled articles added to the existing training set, the precision and recall rates obtained with Rules 2 and 3 are higher than those without using Rule 2 or Rule 3. Therefore, adding unlabeled samples sequentially and updating the word dictionary by putting in the newly pseudo-labeled samples improve the performance of decision-directed approximation. Table 6 indicates that utilizing unlabeled data increases the recall rate. And most of the precision rates are higher than the one obtained without using unlabeled data. This means that the overall performance of our model is better compared to using only the SVM classifier and GN training zones. The recall rate represents false negative errors. In GS zone detection, reducing false negative errors is much more important than reducing false positive errors.

## 6. CONCLUSION

In this paper, we discussed a semi-supervised learning approach for grant support zone detection from online medical journal articles. We adopted an SVM classifier in decision-directed approximation and proposed several rules to improve the performance of our model. Experimental results show that unlabeled data can be utilized to bootstrap grant support zone classification.

We observed that the precision rate drops after 1000 unlabeled samples are classified and added to the training set. Researchers have long realized that training with unlabeled data can degrade classifier performance in some situations [12, 13]. Many semi-supervised methods cannot perform well if the decision boundary falls through dense regions[14]. Possible extension of this work is to use transductive support vector machines to enforce maximum margin on both labeled and unlabeled data.

## 7. ACKNOWLEDGEMENTS

We thank Loc Tran for downloading the journal article HTML files and Dr. Jong Woo Kim for collecting the articles containing grant support information. This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

## REFERENCES

[1] Nigam K., McCallum A.K., Theun S., and Mitchell T.M., "Text classification from labeled and unlabeled documents using EM," Machine Learning 39, 103-134 (2000).

[2] Yarowsky D., "Unsupervised word sense disambiguation rivaling supervised methods," *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 189-196 (1995).

[3] Blum A., Mitchell T., "Combining labeled and unlabeled data with co-training," COLT: *Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann, 92-100 (1998).

[4] Joachims T., “Transductive inference for text classification using support vector machines,” *Proc. of 16th International Conference on Machine Learning*, 200-209 (1999).

[5] Thoma G.R., Le D., Kim I.C., Zou J., Kim J., Tran L.Q., Moon C.W., “Automation to accelerate the production of MEDLINE,” *Technical Report for Board of Scientific Counselors*, Communications Engineering Branch, National Library of Medicine, April 2008.

[6] Zou J., Le D., Thoma G.R., “Combining DOM tree and geometric layout analysis for online medical journal article segmentation,” *Proc. Joint Conference on Digital Libraries*, 119-128 (2006).

[7] Zou J., Le D., Thoma G.R., “Online medical journal article layout analysis,” *Proc. SPIE-IS&T Electronic Imaging 2007, 14th Document Recognition and Retrieval Conference* 6500, 1-12 (2007).

[8] Galavotti L., Sebastiani F., and Simi M., “Experiments on the use of feature selection and negative evidence in automated text categorization,” *Proc. ECDL*, 59-68 (2000).

[9] Sebastiani F., “Machine learning in automated text categorization,” *ACM Computing Surveys* 34(1), 1-47 (2002).

[10] Duda R.O., Hart P. E., and Stork D.G., [Pattern Classification], John Wiley and Sons, New York, 2000.

[11] Chang C.-C. and Lin C.-J., “LIBSVM: a library for support vector machines,” 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[12] Elworthy D., “Does Baum-Welch re-estimation help taggers?” *Proc. of the 4th Conference on Applied Natural Language Processing*, 53-58 (1994).

[13] Cozman F., Cohen I., and Cirelo M., “Semi-supervised learning of mixture models,” *Proc. of the 20th International Conference on Machine Learning*, 99-106 (2003).

[14] Zhu X., “Semi-supervised learning literature survey,” *ICML*-2007 Tutorial, Corvallis, OR, 2007.