2012 11th International Conference on Machine Learning and Applications

Taxonomic Dimensionality Reduction in Bayesian Text Classification

Richard McAllister
Department of Computer Science
Montana State University
Bozeman, MT
Email: mcallis@cs.montana.edu

John Sheppard
Department of Computer Science
Montana State University
Bozeman, MT
Email: john.sheppard@cs.montana.edu

Rafal Angryk
Department of Computer Science
Montana State University
Bozeman, MT
Email: angryk@cs.montana.edu

***Abstract*—Lexical abstraction hierarchies can be leveraged to provide semantic information that characterizes features of text corpora as a whole. This information may be used to determine the classification utility of the dimensions that describe a dataset. This paper presents a new method for preparing a dataset for probabilistic classification by determining, *a priori*, the utility of a very small subset of taxonomically-related dimensions via a Discriminative Multinomial Naïve Bayes process. We show that this method yields significant improvements over both Discriminative Multinomial Naïve Bayes and Bayesian network classifiers alone.**

I. Introduction

Dimensionality is a major factor in text classification (reference the "curse of dimensionality" [1]). Methods that allow data to be represented with lower dimensionality must remove those from consideration those dimensions that contribute less information while retaining those dimensions that are more useful in determining class membership. However, the remaining dimensions reflect a changed representation of the data that they describe.

We present a process for maximizing the benefit that can be obtained by using a taxonomically-driven method of dimensionality reduction. This process applies a Discriminative Multinomial Naïve Bayes (DMNB) process for determining the new probability distribution for the smaller index, and finally, classification using a Bayesian network classifier. We show that this new process is effective at low dimensionality and exceeds the performance of DMNB and the Bayesian network classifier alone.

This paper is organized as follows: In the next section we review work related to text classification and Bayesian networks. This is followed by the definition of several concepts required by our approach. We then provide a detailed account of our approach to text classification, a discussion of the experimental methodology, and the results of our experiments. Lastly, we draw several conclusions from these experiments, review the contributions of this paper, and suggest several areas for future work.

II. Background

A. Bayesian Networks

Bayesian networks are directed acyclic graphs that represent joint probability distributions over a set of random variables [2], [3]. They provide an intuitive and compact representation of the joint distribution over this set of variables by exposing the variables' dependencies. They also provide the means for more efficient statistical inference than working with the full joint distribution.

Each node in a Bayesian network corresponds to a random variable in the domain of interest. Each directed edge in the graph corresponds to a path of influence from a parent node to a child node. In a Bayesian network, each variable x_i is conditionally independent of its non-descendants given its parents. The joint distribution over all random variables $x_i \in X$ in a Bayesian network may be represented as follows.

$$P(X) = P(x_1, \ldots, x_n) = \prod_i P(x_i \mid \text{Pa}(x_i))$$

A Bayesian network may also be used as a classifier. This involves calculating, for each instance, the probability of a particular class variable y given the values of the remaining variables in the network X and returning the class with the highest posterior probability.

$$\text{argmax}_y P(y|X)$$

where

$$P(y \mid X) \propto P(y) \prod_{x \in X} P(x \mid \text{Pa}(x))$$

where y is the class variable and X is the set of non-class variables, i.e. the set of givens [4].

B. Related Work

1) Probabilistic Text Classification: We follow the conditions set forth by Lam and Low [5], who built a Bayesian network text classifier automatically from training text. These conditions are that the edges in the network that exist between a class and a feature must run from the class to the feature.

978-0-7695-4913-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICMLA.2012.93

2) Link-Based Classifiers: Much recent work involving probabilistic graphical treatments of text classification pertains to link analysis, i.e. using the hyperlinks in web documents, for the classification of web pages. Motivated by the then-emerging interest in hypertext mining, Getoor and Lu [6] proposed a framework for modeling link distributions to improve classification accuracy [6]. Popescul, et al. [7] used statistical relational learning to predict where (i.e. in which journals, conference proceedings, etc.) scientific papers will be published. Their approach was based on word counts, citations, co-citations, and word co-occurrences.

3) Taxonomically-Enhanced Data Structures: Caragea, et al. [8] proposed a classifier based on an ontology-extended data structure for classifying semi-structured data. Their ontology essentially consisted of an "Abstraction Hierarchy" that the authors do not specify but define mathematically. We used the WordNet [9] taxonomy for this purpose, without considering links between documents.

Hossain, et al. [10] used WordNet to create "document graphs" which are a form of instantiated taxonomy used for graph-based hierarchical agglomerative clustering. McAllister, et al. [11] [12] used abstraction analysis based on WordNet hypernyms for both information retrieval and text classification.

III. Approach

We apply an abstraction process to reduce dimensionality as a primary step [13]. We then determine the utility of each dimension for classification in each target class for the reduced set of dimensions using DMNBtext [14]. Following this, we perform classification using a Bayesian network classifier.

A. Dimensionality Reduction Strategies

1) Taxonomic Abstraction: Taxonomic Abstraction is the process of determining superordinate and subordinate relationships between words using a predefined hierarchical taxonomy. A word distributes its weight to its each of its parent words in equal proportion, and this distribution continues to the root of the taxonomy. This approach produces instances of a taxonomy that represent a document's footprint in relation to the taxonomy.

For simplicity, we use only nouns, removing from the documents the words that have no noun representation in WordNet. An advantage in this approach is that the entire hierarchy has a single root at the word "entity." Abstraction in this single-rooted hierarchy allows the creation of term-document matrices that are less sparse because all abstractions have to converge at "entity."

2) Latent Semantic Analysis: For comparison, we used Latent Semantic Analysis (LSA), which statistically infers relationships between words in a document corpus. Using the words that have the n greatest variances in their TF×IDF representation for the items in the new index [15], produces an n-rank approximation of a term-document matrix.

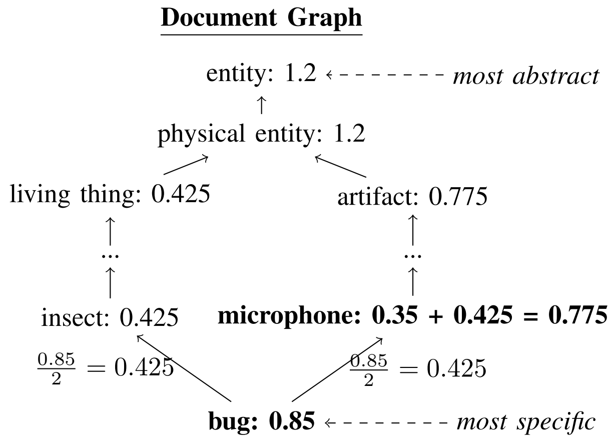

Figure 1. Example of a Document Graph for a document containing two keywords: 'bug' and 'microphone'.

B. System Architecture

1) Term-Document Matrix: We use a data structure called a Term-Document Matrix (TDM) for our classification experiments. The TDM is a matrix M of vectors $\mathbf{m}$ where $m_{i,j}$ is the measure of word relevance of word i to document j. In this case, the TDM is comprised of word vectors that use the Term Frequency × Inverse Document Frequency (TF×IDF) measure [16] as their elements.

2) Document Graphs and Abstraction: For dimensionality reduction through Taxonomic Abstraction we use Document Graphs, which are created by tracing the hypernym superstructures (paths to the root in WordNet) of all nouns in a document. Figure 1 shows a document graph for a hypothetical document containing two words: 'bug' and 'microphone'. There are two different senses of the word 'bug' shown; that of the insect and that of the miniature microphone.

The values in the graph reflect artificial TF×IDF values for the sentence containing the words 'bug' and 'microphone'. Before the document graphing process, the word 'bug' has the TF×IDF value 0.85 and the word 'microphone' has the TF×IDF value of 0.35. Since, in this example, the word 'bug' has two hypernyms, namely 'insect' and 'microphone', 'bug' divides its support (initially the TF×IDF value) evenly among both, adding half of its support (0.425) to each hypernym. Since 'microphone' already has a support of 0.35, its support becomes 0.35 + 0.425 = 0.775 because of the contribution of half of the support of the word 'bug'. The rest of the values represent a propagation of TF×IDF values from hyponym to hypernym, with all values converging at the word 'physical entity' since both 'living thing' and 'artifact' are physical entities.

An Abstraction Path is a path in a document graph from leaf to root or from a location of changing support to root containing all vertices along that path. The weight of this

Abstraction Path is the weight of the vertex most distant from the root. Abstraction Paths are extracted from the document graphs to obtain a profile for the desired level of abstraction. In WordNet, these levels of abstraction yield dimensionalities of 1 at level 1, 3 at level 2, 17 at level 3, and 194 at level 4.

3) Dimensionality Reduction via Latent Semantic Analysis: For dimensionality reduction via latent semantic analysis [15] the same dimensionality that would result from cutting the abstraction paths at levels 1, 2, 3, and 4 was used (1, 3, 17, and 194 dimensions, respectively). This is because a direct comparison of the effectiveness of the number of dimensions can be achieved if the two techniques are compared using the same number of dimensions. Since dimensionality reduction via taxonomic abstraction at, for example, level 2 produces 3 dimensions, the same number of dimensions is selected for the dimensional reduction via latent semantic analysis.

4) Log-Ratio Matrices and Dimension Weighting: From the dimensionality reduction step we obtain a Dimensionally Reduced TDM (DRTDM), which is a TDM in which some dimensions have been deleted or combined as a result of the dimensionality reduction processes, resulting in a TDM of lower dimensionality. We normalize this DRTDM by using each measurement's z-value, or standard score, calculated as

$$z = \frac{x - \mu}{\sigma}$$

where x is the current value being normalized, μ is the mean and σ is the standard deviation.

The Log-Ratio Matrix (LRM) obtained through the use of DMNBtext is a matrix in which each element $l_{i,j}$ is calculated as

$$l_{i,j} = \log\left(\frac{P(t_i \mid c_j)}{P(t_i \mid \neg c_j)}\right)$$

where $P(t_i \mid c_j)$ is the probability that a document contains term t_i given the document is in class c_j and $P(t_i \mid \neg c_j)$ is the probability that a document contains term t_i given the document is not in class c_j. Conceptually, it contains a rating of the usefulness of each dimension for classification in the entry for each class. The LRM is obtained via the DMNBtext classification algorithm [14].

To construct the Bayesian network for final classification, the Weighted, Dimensionally Reduced Term-Document Matrix (WDM) is used. To produce the WDM, each entry in the DRTDM, that entry's class is located in the LRM and each of the items in its term document vector is multiplied by its corresponding weight therein. This is done to reinforce dimensions that are more useful in determining class membership for an entry in the DRTDM and penalize those that are less useful. Then each of the columns in the resulting WDM becomes an example for training the network where the rows correspond to the features (i.e., variables) of the network.

5) Classification via Bayesian Network: The Bayesian networks for classification were constructed using the WDMs created from WordNet's hypernyms for the abstraction experiments and the WDMs created using the lower dimensionality achieved by using LSA for the LSA experiments. In other words, each document's representation was replaced with the representation using this new, lower dimensionality for both abstraction and LSA.

Figure 2 shows the Bayesian network that resulted from the analysis at WordNet's third level. Using hypernyms at this level results in a dimensionality of 17. Each synset in the graph has an edge leading directly from the *class* vertex. This is a result of this network starting as a Naïve Bayes network, with none of the original connections between the class variable and the other variables being severed as the rest of the Bayesian network takes shape.

The Bayesian network in Figure 2 expresses some interesting relationships among hypernyms in the corpus. According to the network, there is a V structure with the word 'set' at the junction. This suggests that the presence of the word 'set' as a hypernym of a word in a document activates the path from 'process' to 'class' through the word 'set' even though 'process' is also directly connected to 'class'. Being siblings in the WordNet hierarchy, 'process' and 'set' have a relationship that does not seem to suggest this new relationship.

IV. Experiments and Evaluation

A. Datasets

We conducted experiments on two datasets: the 20 Newsgroups dataset [17] and the Reuters-21578 dataset [18]. Using the 20 Newsgroups dataset is convenient because it provides a default document labeling via the newsgroups to which each conversation was posted. Also, it provides an interesting example of noisy, unstructured data.

The Reuters-21578 Text Categorization Test Collection [18] is a collection of documents that appeared on the Reuters newswire service in 1987. It provides categories of general subject matter for the documents, which we use as classification labels. This dataset was less intuitive to use for this investigation, as there may be more than one category per document. We handled this using the approach of Qian et al. [19], where documents from the top 10 largest categories of the 'ModApte' split of the dataset are used.

For training and testing, the datasets were divided in half with one half being used to obtain the log-ratio matrix by using the DMNBtext classifier, and the other being used for 10 fold cross-validation with the Bayesian network classifier. We did this to avoid documents that were used to obtain the log-ratio matrix in the final classification experiments.

B. Results

Tables I through IV compare the class-by-class F-measure on the 20 Newsgroups dataset and the Reuters 21578 dataset.

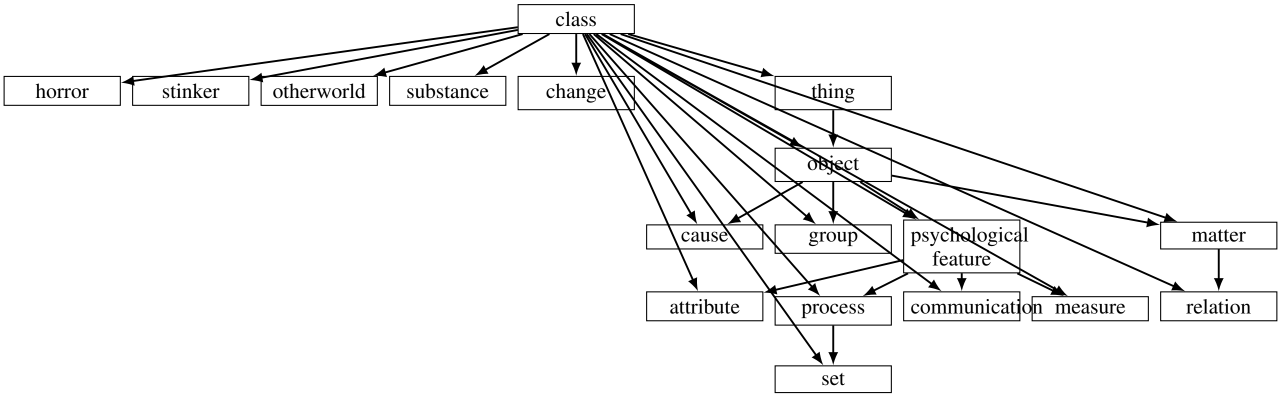

Figure 2. Bayesian network depicting hypernym relationships

20 Newsgroups Dataset				
	Dimensionality			
	1	3	17	194
rec.motorcycles	0.094	0.086	0.107	0.11
comp.sys.mac. hardware	0	0	0.02	0.134
talk.politics.misc	0	0	0	0.064
soc.religion.christian	0.084	0.053	0.129	0.221
comp.graphics	0	0	0.098	0.216
sci.med	0	0.01	0	0.131
talk.religion.misc	0	0	0	0
comp.windows.x	0	0	0.13	0.328
comp.sys.ibm.pc. hardware	0	0	0.004	0.205
talk.politics.guns	0	0	0	0.113
alt.atheism	0	0	0	0.01
comp.os.ms-windows. misc	0	0	0.076	0.213
sci.crypt	0	0.039	0.043	0.256
sci.space	0.069	0.102	0.035	0.355
misc.forsale	0	0	0.142	0.224
rec.sport.hockey	0	0.091	0.071	0.264
rec.sport.baseball	0	0	0.095	0.182
sci.electronics	0	0	0.097	0.153
rec.autos	0.097	0	0.011	0.143
talk.politics.mideast	0	0	0.04	0.274
weighted average:	0.019	0.021	0.058	0.186
percent correctly classified:	5.47	5.60	8.32	20.29
Reuters 21578 Dataset				
earn	0.628	0.614	0.642	0.751
wheat	0	0	0	0
money-fx	0	0	0	0
corn	0	0	0	0
trade	0	0	0	0
acq	0	0	0.197	0.368
grain	0	0	0	0.178
interest	0	0	0	0
crude	0	0	0	0
ship	0	0	0	0
weighted average:	0.287	0.272	0.329	0.43
percent correctly classified:	45.74	44.26	47.02	52.77

Table I
DMNBtext on Abstraction Indexing

Notice that, even at a dimensionality of 1, data model enhancement using the Bayesian network classifier allowed the correct classification of some documents into some classes that had not even been represented without the *a-posteriori* Bayesian network classification. This suggests that the posterior distribution of the classes (after DMNB model construction) affected classification with respect to the entire collection of classifications.

As shown in Tables I and III, the majority of the F-measure values are 0 for a dimensionality of 1. The configuration of entries of 0 is different in both of these tables for a dimensionality of 2. Sometimes, as in the case of the 'rec.autos' row of Table I, the F-measure drops to 0 as the dimensionality increases from 1 to 3. To explain this, the three dimensions represented in the second column do not include the dimension represented in the column corresponding to one dimension. The same is the case for Table III. The dimensions represented in the configuration of three dimensions do not necessarily include those represented in the configuration of one dimension.

Tables II and IV show the results of experiments run using abstraction indexing with the Bayes Net classifier and LSA using the Bayes Net classifier, respectively. Interestingly, the highest percent correctly classified value occurs at 194 dimensions for abstraction indexing and at 3 dimensions for LSA for the 20 Newsgroups dataset. For Reuters 21578 this is different, possibly because this dataset does not deal with classifications that are mutually exclusive. For 20 Newsgroups, however, it suggests that LSA reaches its optimal dimensionality prematurely, before there is a chance for an expressive index to be calculated.

V. Conclusions and Contributions

We draw two major conclusions here. The first is that, provided with an effective, semantically-driven dimensionality reduction technique such as Taxonomic Abstraction, the Discriminative Multinomial Naïve Bayes classifier may be

20 Newsgroups Dataset				
	Dimensionality			
	1	3	17	194
rec.motorcycles	0.551	0.705	0.832	0.98
comp.sys.mac. hardware	0.59	0.579	0.7	0.948
talk.politics.misc	1	1	0.998	0.996
soc.religion.christian	0.984	0.984	0.985	0.98
comp.graphics	0.911	0.911	0.952	0.947
sci.med	0.918	0.941	0.938	0.986
talk.religion.misc	1	1	1	0.998
comp.windows.x	0.583	0.574	0.818	0.962
comp.sys.ibm.pc. hardware	0.683	0.672	0.662	0.971
talk.politics.guns	0.989	0.992	0.991	0.992
alt.atheism	0.996	0.999	0.991	0.985
comp.os.ms-windows. misc	0.751	0.786	0.871	0.952
sci.crypt	0.798	0.899	0.944	0.988
sci.space	0.957	0.972	0.975	0.991
misc.forsale	0.587	0.613	0.774	0.834
rec.sport.hockey	0.793	0.887	0.931	0.967
rec.sport.baseball	0.645	0.764	0.848	0.96
sci.electronics	0.859	0.894	0.896	0.984
rec.autos	0.722	0.773	0.85	0.967
talk.politics.mideast	0.999	0.991	0.998	0.987
weighted average:	0.81	0.84	0.894	0.968
percent correctly classified:	81.10	84.12	89.46	96.64

Reuters 21578 Dataset				
earn	0.995	0.957	0.998	0.99
wheat	0.938	0	0.973	1
money-fx	0.977	0.97	0.958	0.96
corn	0	0.69	0.923	0.97
trade	0.902	0.962	1	1
acq	0.934	0.94	0.979	0.99
grain	0.954	0.947	1	0.986
interest	0.857	0.974	1	0.971
crude	0.955	0.957	1	0.982
ship	0.706	0.96	1	1
weighted average:	0.93	0.909	0.989	0.987
percent correctly classified:	94.26	92.77	98.94	98.72

Table II
DMNBtext Preprocessing Abstraction Indexing With Bayes Net Classifier

20 Newsgroups Dataset				
	Dimensionality			
	1	3	17	194
rec.motorcycles	0	0.074	0.023	0.448
comp.sys.mac. hardware	0	0	0.142	0.367
talk.politics.misc	0	0	0	0.09
soc.religion.christian	0.069	0	0.166	0.479
comp.graphics	0	0	0.123	0.371
sci.med	0	0.023	0	0.218
talk.religion.misc	0	0	0	0
comp.windows.x	0.087	0.166	0.402	0.468
comp.sys.ibm.pc. hardware	0	0.048	0.335	0.332
talk.politics.guns	0	0	0	0.393
alt.atheism	0	0	0	0.07
comp.os.ms-windows. misc	0	0.416	0.398	0.461
sci.crypt	0.087	0.095	0.032	0.647
sci.space	0.084	0.093	0.028	0.378
misc.forsale	0	0.011	0.125	0.337
rec.sport.hockey	0.086	0.122	0.131	0.538
rec.sport.baseball	0.071	0	0.288	0.297
sci.electronics	0	0	0	0.087
rec.autos	0	0	0.131	0.471
talk.politics.mideast	0	0	0	0.53
weighted average:	0.026	0.057	0.124	0.364
percent correctly classified:	5.45	9.53	15.72	36.01

Reuters 21578 Dataset				
earn	0.639	0.592	0.614	0.614
wheat	0	0	0	0
money-fx	0	0	0	0
corn	0	0	0	0
trade	0	0	0	0
acq	0	0	0	0
grain	0	0	0	0
interest	0	0	0	0
crude	0	0	0	0
ship	0	0	0	0
weighted average:	0.3	0.249	0.274	0.274
percent correctly classified:	46.92	42.04	44.16	44.16

Table III
DMNBtext on Latent Semantic Analysis

used effectively to supplement the weighting of a reduced term-document matrix. The second is that, provided with this reduced term-document matrix, Bayesian network classifiers can achieve excellent accuracy at very low dimensionality. Weighting the terms properly, after the initial model construction, can also allow the Bayesian network classifier to classify documents correctly into classes that would have been neglected by the initial classification procedure.

We have also provided two main contributions. The first is a method of enhancing accuracy of text classification where the classification dimensions have been drastically reduced. This is important in that it may save index space and facilitate pruning of the document search space. The second contribution is information about how the accuracy of classification varies as the level of index abstraction varies. This contribution may be seen in Tables I through IV, where the results are broken down by dimensionality.

VI. Future Work

Abstraction and the use of ontologies provides an excellent basis for text classification in information retrieval. Figure 2 shows interesting lexical relationships exposed by the construction of the network. An addition to the WordNet taxonomy that includes these relationships may be instructive and may lead to advances in classification using the enhanced taxonomy as a reference.

We are also interested in the treatment of taxonomies, such as WordNet, as Bayesian networks. In this case, each synset would be treated as a variable and the distributions

20 Newsgroups Dataset				
	Dimensionality			
	1	**3**	**17**	**194**
rec.motorcycles	0.955	0.902	0.947	0.743
comp.sys.mac. hardware	0.82	0.842	0.733	0.613
talk.politics.misc	1	1	0.995	0.986
soc.religion.christian	1	1	0.981	0.95
comp.graphics	0.984	0.984	0.915	0.625
sci.med	0.999	0.996	0.996	0.914
talk.religion.misc	1	1	1	1
comp.windows.x	0.718	0.8	0.851	0.743
comp.sys.ibm.pc. hardware	0.697	0.781	0.711	0.514
talk.politics.guns	1	1	0.993	0.96
alt.atheism	1	1	0.988	0.967
comp.os.ms-windows. misc	0.759	0.899	0.834	0.582
sci.crypt	0.946	0.98	0.881	0.839
sci.space	1	1	0.985	0.916
misc.forsale	0.695	0.742	0.718	0.49
rec.sport.hockey	0.983	0.979	0.817	0.869
rec.sport.baseball	0.964	0.915	0.856	0.853
sci.electronics	0.954	0.989	0.968	0.816
rec.autos	0.872	0.897	0.748	0.649
talk.politics.mideast	1	0.999	0.992	0.975
weighted average:	0.914	0.933	0.892	0.791
percent correctly classified:	91.46	93.30	89.10	78.58
Reuters 21578 Dataset				
earn	0.985	0.966	1	1
wheat	0.952	0.774	1	1
money-fx	0.75	0.895	0.974	0.769
corn	0	0	0.696	0.643
trade	0.87	0.895	0.963	1
acq	0.917	0.937	0.96	0.89
grain	0.944	0.907	0.986	0.983
interest	0.629	0.857	0.889	0.773
crude	0.884	0.898	0.875	0.764
ship	0.9	0.875	0.824	0.8
weighted average:	0.897	0.913	0.963	0.926
percent correctly classified:	91.08	92.36	96.39	92.78

Table IV
DMNBTEXT ON LATENT SEMANTIC ANALYSIS WITH BAYES NET CLASSIFIER

for these variables would be calculated based on the hierarchical dynamics that emerge through the creation of document graphs. Since text processing involves very large and dynamic datasets, exploring dynamic updating for large, constantly changing text corpora would also be useful.

REFERENCES

[1] R. Bellman, *Dynamic Programming*, ser. Dover Books on Mathematics. Dover Publications, 2003. [Online]. Available: http://books.google.com/books?id=fyVtp3EMxasC

[2] J. Pearl, *Probabilistic reasoning in intelligent systems - networks of plausible inference*, ser. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.

[3] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[4] (2011, April) Weka documentation. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/index documentation.html

[5] W. Lam and K.-F. Low, “Automatic document classification based on probabilistic reasoning: model and performance analysis,” in *Systems, Man, and Cybernetics, 1997. 'Computational Cybernetics and Simulation'., 1997 IEEE International Conference on*, vol. 3, 1997, pp. 2719–2723 vol.3.

[6] Q. Lu and L. Getoor, “Link-based text classification,” in *IJCAI Workshop on Text Mining and Link Analysis*, 2003.

[7] A. Popescul, L. H. Ungar, S. Lawrence, and D. M. Pennock, “Statistical relational learning for document mining,” in *Proceedings of the Third International Conference on Data Mining*, 2003, pp. 275–282.

[8] C. Caragea, D. Caragea, and V. Honavar, “Learning link-based classifiers from ontology-extended textual data,” in *IEEE International Conference On Tools With Artificial Intelligence*. IEEE Computer Society, 2009, pp. 354–361.

[9] (2011, December) Wordnet: a lexical database for the english language. [Online]. Available: http://wordnet.princeton.edu/

[10] M. S. Hossain and R. A. Angryk, “Gdclust: A graph-based document clustering technique.” in *ICDM Workshops*. IEEE Computer Society, 2007, pp. 417–422.

[11] R. A. McAllister and R. A. Angryk, “An efficient abstraction-based data model for information retrieval,” in *The 22nd Australasian Joint Conference on Artificial Intelligence*, 2009, pp. 567–576.

[12] ——, “Evaluation of abstraction-based data models for text via supervised learning methods,” in *SIAM International Conference on Data Mining*, 2010.

[13] ——, “Abstracting for dimensionality reduction in text classification,” *to appear in International Journal of Intelligent Systems*, 2013.

[14] J. Su, H. Zhang, C. Ling, and S. Matwin, “Discriminative parameter learning for bayesian networks,” in *ICML 2008'*, 2008.

[15] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, “Indexing by latent semantic analysis,” *Jounal of the American Society for Information Science*, vol. 41-6, pp. 391–407, 1990.

[16] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, July 2008.

[17] (2009, May) Home page for 20 newsgroups data set. [Online]. Available: http://people.csail.mit.edu/jrennie/20Newsgroups/

[18] (2011, December) Reuters-21578 text categorization test collection. [Online]. Available: http://www.daviddlewis.com/resources/testcollections/reuters21578/

[19] T. Qian, H. Xiong, Y. Wang, and E. Chen, “On the strength of hyperclique patterns for text categorization ,” *Jounal of Information Science*, 2007.