Pruning the vocabulary for better context recognition

Madsen, Rasmus Elsborg; Sigurdsson, Sigurdur; Hansen, Lars Kai; Larsen, Jan

Published in:
Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.

Link to article, DOI:
10.1109/ICPR.2004.1334270

Publication date:
2004

Document Version
Publisher's PDF, also known as Version of record

Pruning The Vocabulary For Better Context Recognition

Rasmus Elsborg Madsen, Sigurdur Sigurdsson, Lars Kai Hansen and Jan Larsen
Informatics and Mathematical Modelling, Technical University of Denmark
Richard Petersens Plads, Building 321, DK-2800 Kongens Lyngby, Denmark
web: www.imm.dtu.dk, e-mail: rem,siggi,lkh,jl@imm.dtu.dk

***Abstract*— Language independent 'bag-of-words' representations are surprisingly effective for text classification. The representation is high dimensional though, containing many non-consistent words for text categorization. These non-consistent words result in reduced generalization performance of subsequent classifiers, e.g., from ill-posed principal component transformations. In this communication our aim is to study the effect of reducing the least relevant words from the bag-of-words representation. We consider a new approach, using neural network based sensitivity maps and information gain for determination of term relevancy, when pruning the vocabularies. With reduced vocabularies documents are classified using a latent semantic indexing representation and a probabilistic neural network classifier. Reducing the bag-of-words vocabularies with 90%-98%, we find consistent classification improvement using two mid size data-sets. We also study the applicability of information gain and sensitivity maps for automated keyword generation.**

I. Introduction

The world wide web is an unstructured and fast growing database. Today's search tools often leave web users in frustration by the low precision and recall [6]. It is widely believed that machine learning techniques can come to play an important role in web search. Ambitious plans have been launched for supporting intelligent use of the web, i.e., a "semantic web" [4]. IBM's WebFountain [5] and the Stanford University semantic web platform TAP [13] are examples of machine learning methods coming into play, making human web navigation easier. Here we consider web content mining in the form of internet document classification - an information retrieval (IR) aspect of web-mining [17]. Internet documents contain text, hyper-links, meta-data, images, and other multimedia content which can be used for classification [17], [16]. This paper focuses on classification based on text part, i.e., text categorization. Text categorization is the process of creating a supervised automatic text classifier, by means of machine learning techniques. The classifier labels documents from the corpus $\mathcal{D} = [d_1, \cdot, d_j, \cdot, d_{|\mathcal{D}|}]$ into a set of classes $\mathcal{C} = [c_1, \cdot, c_k, \cdot, c_{|\mathcal{C}|}]$, based on an initial set of labeled documents.

Generic text categorization systems are based on the bag-of-words representation, which is surprisingly effective for the task. In the bag-of-words representation we summarize documents by their term histograms. The main motivation for this reduction (removing the semantics) is that it is easily automated and needs minimal user intervention beyond filtering of the term list. The term list typically contains in the range of $10^3 - 10^5$ terms, hence further reduction is necessary for most pattern recognition devices. Latent semantic indexing (LSI) [12], [11] aka principal component analysis is often used to construct low dimensional representations. LSI is furthermore believed to reduce synonymy and polysemy problems [11], [19]. Synonymy is when multiple words have the same meaning and polysemy is when a single word have multiple meanings. Although LSI and other more elaborate vector space models have been successful in text classification in small and medium size databases, see e.g., [16], [14], it is still not at human level text classification performance. When training classifiers on relatively small databases generalizability is a key issue. How well does a model adapted on one set of data predict the labels of another test data set? Generalizability is in general a function of the number of training cases and of the effective model dimension.

Various methods and techniques have been purposed to improve generalization in text categorization. WordNet [2], a lexical database containing synonym sets and other lexical concept, has been used for classification improvement. In [9], the synonymy part in WordNet has been used to expand term-lists for each text category, enhancing the accuracy of the text classifier significantly. In [15], text classification based WordNet's word meanings has been attempted. These experiments have not given any significant classification accuracy enhancement. On the other hand, the use of words part-of-speech (POS) has showed to improve text categorization generalization. A POS-tagger analyzes sentences and tag words with their part-of-speech, i.e., noun, verb, adverb, number, punctuation, etc. In [3], words have been tagged with their POS, avoiding confusion between similar words with different meanings. This approach resulted in a positive effect on classification accuracy. In [1], a POS-tagger has been used to extract more than 3.000.000 compound words from texts, improving classification accuracy. Using unlabeled documents when categorizing texts has an improving effect. The unlabeled documents can be used in various ways, see e.g. [24], [20] and [31]. In [18], multiple classifiers are combined, and a consensus voting scheme among the classifiers performs better than any single classifier.

In this communication, the aim is to improve generalizability of the supervised document classifier by pruning the document vocabulary $\mathcal{T} = [t_1, \cdot, t_i, \cdot, t_{|\mathcal{T}|}]$, i.e., removing the term which is least suited for discrimination. Many terms

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

posses little or no generalizable discriminative power, and should be regarded as noise. Pruning the vocabulary, the task is to determine the least discriminative terms, based on the training set only. Automated vocabulary reduction has been attempted with success previously in [30]. We here use another method for term reduction, and experiment within the LSI representation. To estimate term relevance we will use Information Gain and scaled sensitivity, which is computed using the so-called NPAIRS split-half re-sampling procedure [29]. Our hypothesis is that sensitivity maps can determine which terms are consistently important, hence, likely to be of general use for classification relative to terms that are of low or highly variable sensitivity.

The rest of this article is organized as follows. In Section II, we discuss the generic bag-of-words approach for text categorization, and the vocabulary pruning methods. In Section III, explains the data sets used for the experiments. Section IV presents the results obtained using vocabulary pruning. Section V concludes on the methods and results.

II. METHODS

Using the generic bag-of-words approach, documents are arranged in a term-document matrix $\mathbf{X}$, where $X_{i,j}$ is the number of times term i occur in document j. The dimensionality of $\mathbf{X}$ is reduced by filtering and stemming. Stemming refers to a process in which words with different endings are merged, e.g., 'train', 'trained' and 'training' are merged into the common stem 'train'. This example also indicates the main problem with stemming, namely that it introduces an artificial increased polysemy. We have decided to 'live with this problem' since without stemming vocabularies would grow prohibitively large. About 500 common non-discriminative stop-words, i.e. ('a', 'i', 'and', 'an', 'as', 'at') are removed from the term list. In addition high and low frequency words are also removed from the term list. The term-document matrix can be normalized in various ways. In [10] experiments with different term weighting schemes are carried out. The term frequency / inverse document frequency (TFIDF) weighting is consistently good among term weighting methods purposed, and is the method generally used. After TFIDF normalization the resulting elements in $\mathbf{X}$ becomes

$$X_{i,j}^{\text{tfidf}} = X_{i,j}^{\text{tf}} \log \frac{|\mathcal{D}|}{DF_i} \tag{1}$$

where DF_i is the document frequency of term i and $X_{i,j}^{\text{tf}}$ is the log normalized term frequency.

$$X_{i,j}^{\text{tf}} = \begin{cases} 1 + \log(X_{i,j}) & \text{if} X_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The length of the documents is often a good prior for predicting the content within a little corpora. While document length might be a solid variable within the corpora, it is likely that this is not generally a valid parameter. The length of the documents is usually normalized to prevent the influence the document length might have. The Frobenius norm is used to length normalize the term document matrix to one.

$$X_{i,j}^{\text{n2tfidf}} = \frac{X_{i,j}^{\text{tfidf}}}{\sqrt{|\mathcal{T}|^{-1} \sum_{i'=1}^{|\mathcal{T}|} X_{i',j}^{\text{tfidf}^2}}} \tag{3}$$

To emphasize the influence of document lengths, the distribution of the term standard deviations for the spam and non-spam documents, in the email data-set, are illustrated in Figure 1,

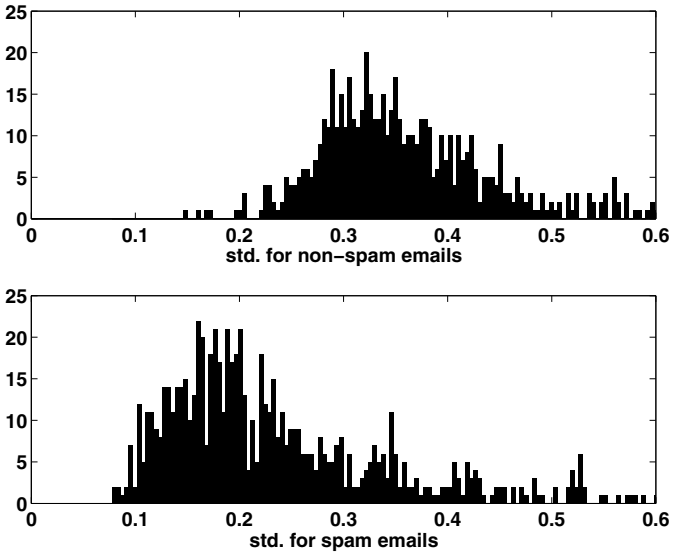

Fig. 1. Distribution of the standard deviation for the email data-set. The distribution for the spam class and the non-spam class varies a lot. The standard deviation is a good discriminator, but probably not general outside this data-set. Using only the standard deviation for classification, the generalization error is 22%.

Using only the standard deviation measure for classification, 78% of the documents can be classified correctly. This clearly shows that document length is a good prior.

It is suggested to use a reduced normalized vocabulary, using sensitivity maps and information gain. The reduction factor ξ determines the fraction of the vocabulary, which is removed.

$$\xi = \frac{|\mathcal{T}| - |\mathcal{T}'|}{|\mathcal{T}|} \tag{4}$$

Where $\mathcal{T}'$ is the new vocabulary, a subset of the full vocabulary $\mathcal{T}$.

Using sensitivity maps for pruning, we use the definition of class specific sensitivity proposed in [32], [28] for a set of N samples,

$$\mathbf{s}_k = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{\partial P(c_k|\mathbf{f}_n)}{\partial \mathbf{x}} \right| \tag{5}$$

and where $P(c_k|\mathbf{f}_n)$ is the posterior probability of class k given the feature vector $\mathbf{f}_n$. $\mathbf{s}_k$ is the K-dimensional sensitivity vector for class k. The K-dimensional derivative is obtained using the projection (8) [28]. A split-half re-sampling procedure is invoked to determine the statistical significance of the

sensitivity [29]. Multiple splits are generated of the original training set and classifiers trained on each of the splits. For each classifier a sensitivity map is computed. Since the two maps obtained from a given split are exchangeable the mean map is an unbiased estimate of the 'true' sensitivity map, while the squared difference is a noisy, but unbiased estimate of the variance of the sensitivity map. By repeated re-sampling and averaging the sensitivity map and its variance are estimated. We finally obtain a scaled sensitivity map by normalization trough the standard deviation.

The scaled sensitivity way of pruning will be compared with information gain pruning. The information gain [30] for the term t_i is defined as:

$$\begin{aligned} IG_{t_i} = & -\sum_{k=1}^{|\mathcal{C}|} P(c_k) \log P(c_k) \\ & + P(t_i) \sum_{k=1}^{|\mathcal{C}|} P(c_k|t_i) \log P(c_k|t_i) \\ & + P(\overline{t}_i) \sum_{k=1}^{|\mathcal{C}|} P(c_k|\overline{t}_i) \log P(c_k|\overline{t}_i), \end{aligned} \tag{6}$$

where $P(t_i)$ is the probability that term t_i appears at least once in a document and $P(\overline{t}_i)$ is the probability that the term does not appear in a document.

The normalized and pruned term document matrix $\mathbf{X}_\mathrm{p}$ is reduced to a feature-document matrix using PCA, carried out by an 'economy size' singular value decomposition,

$$\mathbf{X}_\mathrm{p} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T. \tag{7}$$

Where the orthogonal $|\mathcal{T}'| \times |\mathcal{D}|$ matrix $\mathbf{U}$ contains the eigenvectors corresponding to the non-zero eigenvalues of the symmetric matrix $\mathbf{X}_\mathrm{p}\mathbf{X}_\mathrm{p}^T$. $\mathbf{\Lambda}$ is a $|\mathcal{D}| \times |\mathcal{D}|$ diagonal matrix of singular values ranked in decreasing order and the $|\mathcal{D}| \times |\mathcal{D}|$ matrix $\mathbf{V}^T$ contains eigenvectors of the symmetric matrix $\mathbf{X}_\mathrm{p}^T\mathbf{X}_\mathrm{p}$. The LSI representation is obtained by projecting document histograms on the basis vectors in $\mathbf{U}$,

$$\mathbf{F} = \mathbf{U}^T\mathbf{X} = \mathbf{\Lambda}\mathbf{V}^T. \tag{8}$$

Typically, the majority of the singular values are small and can regarded as noise. Consequently, only a subset of K ($K < |\mathcal{T}'|$) features is retained as input to the classification algorithm. The representational potential of these LSI features is illustrated in Figure 2.

A wide variety of classification algorithms have been applied to the text categorization problem, see e.g., [17]. We have extensive experience with probabilistic neural network classifiers and a well tested ANN toolbox is available [26]. The toolbox adapts the network weights and tunes complexity by adaptive regularization and outlier detection using the Bayesian ML-II framework, hence, requires minimal user intervention [27].

III. Data

Two data-sets, 'Email' [23] and 'WebKB' [7] are used to illustrate and test the hypothesis. No less than ten split-half re-samples are used in all experiments. The Email data-set consists of texts from 1431 emails in three categories: conference (370), job (272) and spam (789). The WebKB set

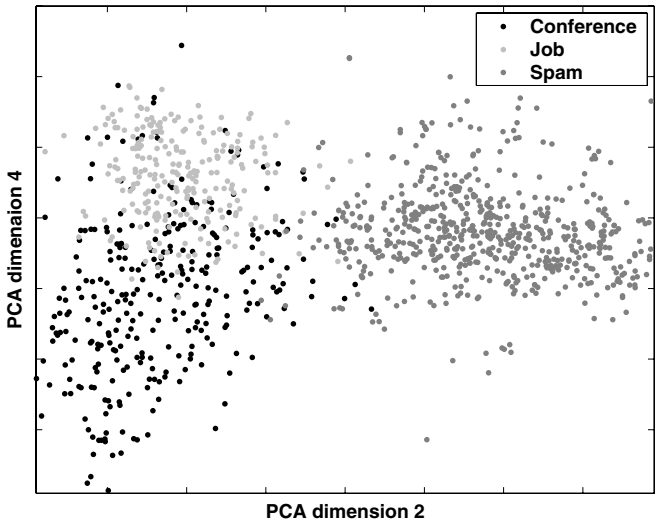

Fig. 2. Illustration of the document distribution in feature space. Here we show the Email corpus projected onto the 2nd and 4th principal directions. In this projection the 'spam' class is well separated while the two other classes in the set ('conferences' and 'jobs') show some overlap.

contains 8282 web-pages from US university computer science departments. Here we have used a subset [8] of 2240 pages from the WebKB earlier used in [14] and [19]. The WebKB categories are: project (353), faculty (483), course (553) and student (851). All html tags were removed from the data-set.

IV. Results

The standard performance measure for text categorization systems is precision and recall. Precision measures how many of the retrieved entries are relevant precision $=$ true positive$/$(true positive $+$ false positive). Recall measures how many relevant entries were found compared to the amount of relevant entries in the collection recall $=$ true positive$/$(true positive $+$ false negative. The F_β measure [21] weights the importance of precision and recall, where $\beta = 1$ weights precision and recall equally,

$$F_\beta = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} \tag{9}$$

Micro averaging [25] over all the classes $|\mathcal{C}|$, is rewarded when classifiers of frequent categories performs well. When each document belongs to less than two classes, which is the case for the collections considered here, the micro averaged precision and recall simplifies to the fraction of correct classified documents. It follows from that the F_1 measure also becomes the fraction of correct classified documents. In the following we use the error function defined as $1 - F_1$.

Preprocessing the documents, all letters in all the terms have been converted to lowercase and punctuations have been removed. A simple stemmer has transformed words with basic endings into their common stem. Preliminary experiments indicated that a reduced feature space of $K = 48$ projections and a neural network classifier with five hidden units were sufficient for the task (data not shown). All results have been

validated using 10 fold split half re-sampling cross validation. The neural network based term sensitivity is a function of the given training set. Terms for which the sensitivity is high but also highly variable are less likely to support generalizability compared to terms that have a consistent high or medium sensitivity. The empirical distribution of mean and standard deviations the terms sensitivities of the Email set are shown in Figure 3.

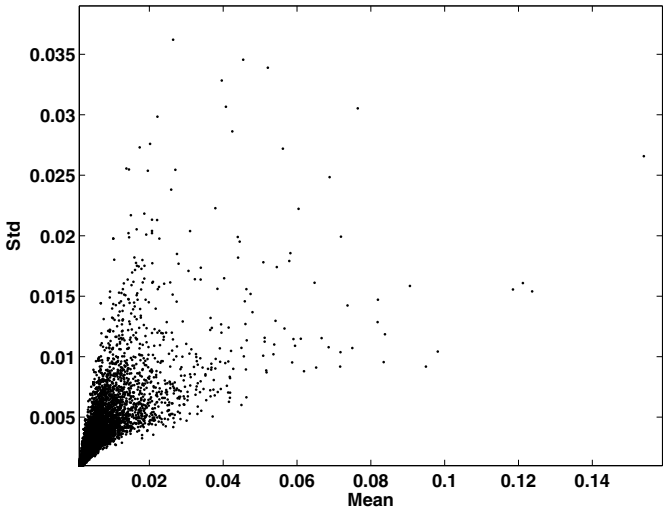

Fig. 3. Mean and standard deviation of the term sensitivity. The most relevant terms have consistently high sensitivity in each re-sampling split, i.e., a high mean and relatively low standard deviation. These terms occupy the lower right part of the plot.

The empirical scaled sensitivities $Z_i = \mu_i/\sigma_i$ of the terms t_i were used to determine the term relevance. Term relevance were also determined using information gain. The two methods are quite different, and so is the distribution of their estimated term relevance. In Figure 4, the distribution of term relevance is shown for the Email data, using the two methods.

Both relevance measure distributions have large slender tails, showing that few terms posses much information. It is likely that the vocabularies can be pruned intensively.

Based on the scaled sensitivities, relevant keywords for the text categories has been extracted. For the Email data the five highest scores for the Conference category are (*Paper, Conference, Deadline, Neural, Topic*) and for the Job category (*Research, Position, Candidate, University, Edt*) and for the Spam category (*Money, Remove, Free, Thousand, Simply*). Similarly, information gain has been used to determine relevant keywords for the Email data. The five highest scores for the Conference category are (*Neural, Conference, Paper, Science, Workshop*) and for the Job category (*University, Research, Candidate, Computational, Position*) and for the Spam category (*Money, Free, Remove, Business, Simply*). The two sets of keywords posses high relevancy for the three classes, though the two methods does not find the exact same keywords.

The vocabularies of the WebKB and Email data are pruned with an increasing reduction factor ξ. The generalization error as function of $1 - \xi$ is shown in Figure 5.

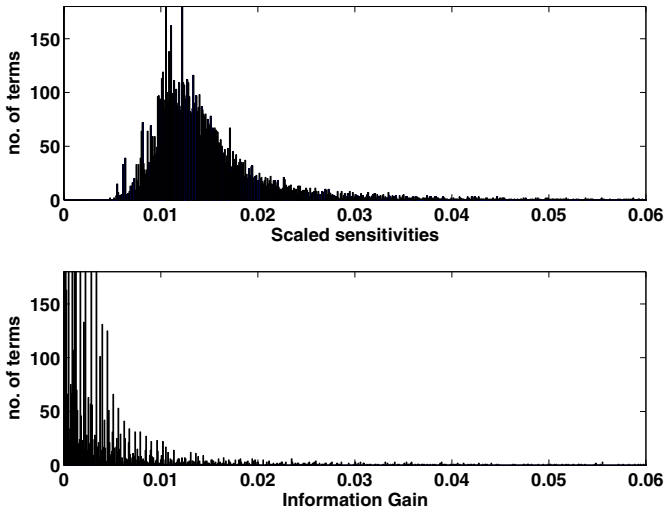

Fig. 4. Distribution of term relevance scores, using scaled sensitivities and information gain, for the Email data. Both distributions have large slender tails, indicating that few terms posses much higher relevancy than others, and intensive pruning should be performed. 10% of the terms have a scaled sensitivity higher than 0.025 and 10% of the terms have information gain higher than 0.008.

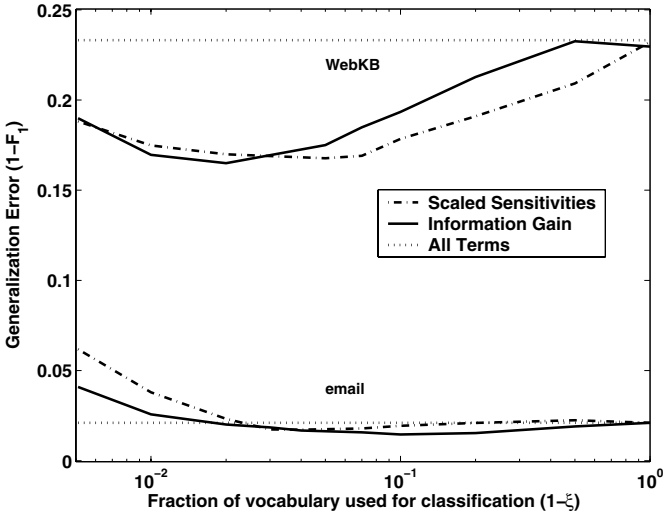

Fig. 5. Generalization error pruning the vocabulary for the Email and WebKB data, using scaled sensitivities and information gain as term relevance measure. Pruning with use of information gain gives slightly better generalization error than when using scaled sensitivities. Reducing the vocabulary with 90%, using Information gain, is optimal for the Email data-set. The generalization error is then reduced with 26%. For the WebKB data-set the lowest generalization error is found, reducing the vocabulary with 98%, where the error is reduced with 29%. The results were found using 20% of the samples for training.

Using all the terms, the generalization classification error rate is 23.3% in the WebKB and 2.1% in Email data. Removing respectively 98% and 90% of the vocabularies with the lowest information gain, the generalization error for the WebKB is reduced to 16.5% and to 1.5% for the Email data. Removing terms with the lowest information gain, the performance is slightly better than when using scaled sensitivities for term removal.

In Figure 6 we show that learning curves are consistently improved for a range of training sets for the WebKB and the Email data, based on a fixed reduction of respectively 98% and 90% of their original vocabulary.

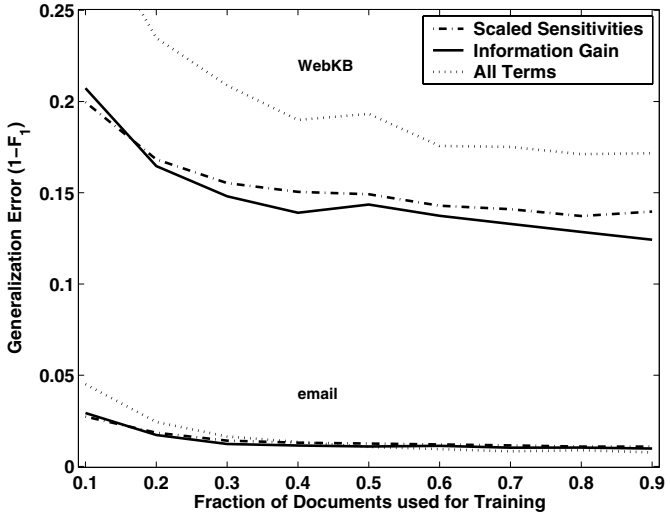

Fig. 6. Learning curves using full and pruned vocabularies. Learning curves shows decreased generalization error for a range of training set sizes. For the WebKB, both pruning methods shows consistent reduced generalization error of about 25% for the whole range of training set sizes. For the Email data, pruning decreases the generalization error when using less than 40% of the data-set for training. When 40% or more of the data-set samples are used for training, the generalization error is not reduced further. Noise within the data might prevent classification from further optimization.

Pruning the vocabulary to a small fraction of the original sizes, results in better generalization in the whole range of training-set sizes, however, a somewhat larger effect for small training sets. For both data-sets, pruning lowers the generalization error with approximately 25%. For the Email set, generalization error is not lowered using 40% of the data or more for training. It is likely that noise within the data-set prevents the classifier from lowering the generalization error any further. For both data-sets information gain is generally slightly better than scaled sensitivities, at determining which terms are relevant for classification. Information gain is significantly cheaper to compute than the scaled sensitivities, which make them the obvious choice among the two methods.

V. Conclusion

Neural network sensitivity maps were introduced in a LSI based context recognition framework. Scaled sensitivity information gain were compared for vocabulary pruning. Using two mid-size data-sets, both methods have consistently shown reduction in text classification error when pruning the vocabularies. Both methods lower the generalization error by approximately 25% over a range of training set sizes. Information gain is generally better at determining the relevant vocabulary information, resulting in slightly better generalization error relative to using scaled sensitivities. Finally, we noted that information gain and the scaled sensitivity are also useful for identifying class specific keywords.

Acknowledgment

The work is supported by the European Commission through the sixth framework IST Network of Excellence: Pattern Analysis, Statistical Modelling and Computational Learning (PASCAL), contract no. 506778.

References

[1] A. Aizawa. Linguistic techniques to improve the performance of automatic text categorization. In *Proceedings of NLPRS-01, 6th Natural Language Processing Pacific Rim Symposium*, pages 307–314, Tokyo, JP, 2001.

[2] Cognitive Science Laboratory at Princeton University. Wordnet 2.0. http://www.cogsci.princeton.edu/ wn/, 2003.

[3] R. Basili, A. Moschitti, and M.T. Pazienza. NLP-driven IR: Evaluating performances over a text classification task. In Bernhard Nebel, editor, *Proceeding of IJCAI-01, 17th International Joint Conference on Artificial Intelligence*, pages 1286–1291, Seattle, US, 2001.

[4] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 2001.

[5] IBM Almaden Research Center. The WebFountain. http://www.almaden.ibm.com/webfountain/publications/.

[6] S. Chakrabarti. Data mining for hypertext: a tutorial survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, 1:1–11, 2000.

[7] CMU-WebKB. The 4 universities data set. http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/, 1997.

[8] CMU-WebKB-2240. A subset of the webkb. http://www.imm.dtu.dk/∼rem/, 1999.

[9] M. De Buenaga Rodríguez, José María Gómez-Hidalgo, and Belén Díaz-Agudo. Using WordNet to complement training information in text categorization. In Ruslan Milkov, Nicolas Nicolov, and Nilokai Nikolov, editors, *Proceedings of RANLP-97, 2nd International Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, BL, 1997.

[10] F. Debole and F. Sebastiani. Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, pages 784–788, Melbourne, US, 2003. ACM Press, New York, US.

[11] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41:391–407, 1990.

[12] G.W. Furnas, S. Deerwester, S.T. Dumais, T.K. Landauer, R.A. Harshman, L.A. Streeter, and K.E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *The 11th International Conference on Research and Development in Information Retrieval*, pages 465–480, Grenoble, France, 1988. ACM Press.

[13] R. Guha and R. McCool. Tap: A semantic web platform. *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 42:557–577, 2003.

[14] L.K. Hansen, S. Sigurdsson, T. Kolenda, F.A. Nielsen, U. Kjems, and J. Larsen. Modeling text with generalizable gaussian mixtures. In *International Conference on Acoustics, Speech and Signal Processing*, pages 3494–3497. IEEE, 2000.

[15] Athanasios Kehagias, Vassilios Petridis, Vassilis G. Kaburlasos, and Pavlina Fragkou. A comparison of word- and sense-based text categorization using several classification algorithms. *Journal of Intelligent Information Systems*, 21(3):227–247, 2003.

[16] T. Kolenda, L.K. Hansen, J. Larsen, and O. Winther. Independent component analysis for understanding multimedia content. In S. Bengio J. Larsen H. Bourlard, T. Adali and S. Douglas, editors, *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, pages 757–766, Piscataway, New Jersey, 2002. IEEE Press.

[17] R. Kosala and H. Blockeel. Web mining research: A survey. In *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, pages 1–15. ACM Press, 2000.

[18] L.S. Larkey and W.B. Croft. Combining classifiers in text categorization. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, pages 289–297, Zürich, CH, 1996. ACM Press, New York, US.

[19] J. Larsen, L.K. Hansen, A.S. Have, T. Christiansen, and T. Kolenda. Webmining: learning from the world wide web. *Computational Statistics and Data Analysis*, 38:517–532, 2002.

[20] J. Larsen, A. S. Have, and Hansen L. K. Probabilistic hierarchical clustering with labeled and unlabeled data. *International Journal of Knowledge-Based Intelligent Engineering Systems*, 6:56–62, 2002.

[21] D.D. Lewis and W.A. Gale. A sequential algorithm for training text classifiers. In W. Bruce Croft and Cornelis J. van Rijsbergen, editors, *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, pages 3–12, Dublin, IE, 1994. Springer Verlag, Heidelberg, DE.

[22] R.E. Madsen and L.K. Hansen. Part-of-speech enhanced context recognition. In *The 17th international conference on pattern recognition*, Cambridge, UK, 2004.

[23] F.Å. Nielsen. Email data-set. http://www.imm.dtu.dk/∼rem/, 2001.

[24] K. Nigam, A.K. McCallum, S. Thrun, and T.M. Mitchell. Learning to classify text from labeled and unlabeled documents. In *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence*, pages 792–799, Madison, US, 1998. AAAI Press, Menlo Park, US.

[25] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34:1–47, 2002.

[26] S. Sigurdsson. The dtu: Artificial neural network toolbox. http://mole.imm.dtu.dk/toolbox/ann/, 2002.

[27] S. Sigurdsson, J. Larsen, L.K. Hansen, P. A. Philipsen, and H. C. Wulf. Outlier estimation and detection: Application to skin lesion classification. In *International conference on acoustics, speech and signal processing*, pages 1049–1052, 2002.

[28] S. Sigurdsson, P.A. Philipsen, L.K. Hansen, J. Larsen, M. Gniadecka, and H.C. Wulf. Detection of skin cancer by classification of Raman spectra. *Accepted for IEEE Transactions on Biomedical Engineering*, 2004.

[29] S.C. Strother, J Anderson, L.K. Hansen, U. Kjems, R. Kustra, J. Siditis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg. The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *Neuroimage*, 15:747–771, 2002.

[30] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.

[31] S. Zelikovitz and H. Hirsh. Using LSI for text classification in the presence of background text. In Henrique Paques, Ling Liu, and David Grossman, editors, *Proceedings of CIKM-01, 10th ACM International Conference on Information and Knowledge Management*, pages 113–118, Atlanta, US, 2001. ACM Press, New York, US.

[32] J.M. Zurada, A. Malinowski, and Cloete I. Sensitivity analysis for minimization of input data dimension for feedforward neural network. In *Proceedings of the IEEE Symposium on Circuits and Systems*, pages 447–450, 1994.