International Journal of Information Processing, 6(4), 42-47, 2012
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Analysis and Experimental Results of Modified Gram-Schmidt Process based Document Classifier

Sumanta Guha [a] and Ananta Raj Lamichhane[b]

[a]Professor, Department of Computer Science and Engineering, Asian Institute of Technology, Pathumthani, Thailand, Contact: guha@ait.ac.th

[b]Department of Computer Science and Engineering, Asian Institute of Technology, Pathumthani, Thailand, Contact: anantalamichhane1@gmail.com

This paper proposes a modified Gram-Schmidt algorithm of a document vector space in order to do classification. The performance of the proposed algorithm was evaluated by comparing it, with commonly known algorithms such as the centroid-based algorithm and latent semantic indexing. Further, to measure efficiency, a modified Gram-Schmidt training set is applied in the centroid-based algorithm and latent semantic indexing as well. Performance measurement was based on two different parameters, *viz*., classification accuracy and closeness of similarity with the corresponding training set. The results shows that modified Gram-Schmidt algorithm is indeed an effective method for dimension reduction prior classification. Moreover, it is easy to code and computationally inexpensive.

Keywords: Centroid, Dimension Reduction, Document Classification, Gram-Schmidt, LSI, Vector Space Model.

1. INTRODUCTION

A problem which faces researchers nowadays is the huge volume of papers and reports available digitally. A researcher may have to shift through hundreds of documents in order to find just a few relevant ones. A first step in the search, typically, is to classify into broad categories the documents returned by the initial query. Automatic document classification is the task of assigning text documents to prespecified categories [1]. Generally, it is defined as content-based assignment to one or more of such categories. Automatic classifiers have been developed using statistical pattern recognition, neural network and machine learning approaches [2]. Automatic classification helps researchers, scientists and students to find required documents. It is vital as well in organizing information in giant digital banks, *e.g.*, online libraries. Retrieving, categorizing, routing and filtering of the documents are all based on text classification.

In document categorization, typically, we already have human indexed training data at hand. A classifier is used to automatically determine to which class a new document should be added [3]. The efficiency and quality of document classification clearly depends on the representation of documents [4]. Generally, classification involves phases such as document indexing, dimensionality reduction, classifier learning, classification and evaluation [5].

Automatic document classication can be a key component of storage and retrieval operations, which are central in databases and data mining. Moreover, in handling massive amounts of data, computational efficiency is crucial. At present various algorithms has been proposed for document classification. The vector space model [6] which represents documents as vectors has been used most widely. Centroid [7] classification techniques use the vector space model to classify the documents. Classification by dimension reduction in the vector space model is an important concept. Some algorithms that perform dimension reduction are Latent semantic analysis (LSI) [8],

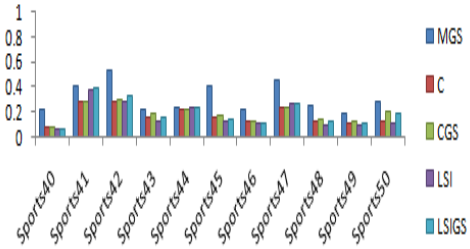

Figure 5. Cosine Similarity Score given by Each Document Treated with their Corresponding Training Set on Different Classifiers.

Table 2
The Percentage Classification Accuracy Achieved by the Different Classification Algorithms.

	MGS	C	CMGS	LSI	LSIMGS
Sport	80	85	85	85	85
Food	90	90	90	90	90
Health	60	70	70	80	80
Religion	90	90	90	90	95
Average	80	83.75	83.75	86.25	87.5

formance of MGS suggests that it employs a sound underlying classification model. The algorithm makes an effort to find out the low-dimensional subspace of the original span of the training set-obtained through a modified Gram-Schmidt process. Data in Table 1 shows that the dimension has been substantially reduced. All vector of the original vector space make angle less than alpha with this reduced dimension space. Reduced parameters in the algorithm are obtained by writing the projected vectors in the new reduced subspace in terms of a basis of that reduced subspace. Thus data are represented in the reduced dimensional space and hence computational efficiency has been achieved.

5. CONCLUSIONS AND FUTURE WORK

This research paper focuses on data representation in a reduced space created using modified a Gram-Schmidt process. The experimental evaluation shows that a modified Gram-Schmidt process based classification algorithm performs well on multiple ranges of data sets. It also shows that the strength of this classifier is its consideration of a threshold angle between document vectors and its projection on the original space to form a reduced space.

The centroid algorithm doesn't perform dimension reduction. The LSI algorithm does, but has lot of computational overhead during dimension reduction [8]. However MGS is efficient and less computationally intensive. As the test range for our experiments is limited, it is not possible to conclusively compare the performance of MGS versus LSI. However it can be claimed that our work shows a new and plausible approach to dimension reduction.

To further prove this new technique as competitive with established methods, more work is needed. Sound experiments of the proposed technique with other established methods should be a goal for future work. Also as all the document vectors of the original vector space make angle less than α – which is some arbitrary user-decided value with reduced dimension space, a theoretical question is to be able to determine some optimality condition on α.

REFERENCES

1. Torkkola K. Linear Discriminant Analysis in Document Classification, *Proceedinds of IEEE ICDM Workshop Text Mining*, 2001.
2. Goller C, Lsoning J, Will T and Wolf W. Automatic Document Classifcation: A Thorough Evaluation of Various Methods, *Machine Learning*, 1:1-11, 2000.
3. Ye J Li, Q Xiong H, Park H, Janardan R and Kumar V. IDR/qr: An Incremental Dimension Reduction Algorithm *via* QR Decomposition, *IEEE Transactions on Knowledge and*

Data Engineering, 17(9):1208-1222, September 2005.
4. Biro I. Document Classification with Latent Dirichlet Allocation, *Unpublished Doctoral Dissertation*, Eotvos Lorand University, 2009.
5. Li J and Sun M. Scalable Term Selection for Text Categorization, *Computational Linguistics*, pages 774-782, June 2007.
6. Salton G, Wong A and Yang C S. A Vector Space Model for Automatic Indexing, *Communications of the ACM*, 18(00):613-620, Nov 1975.
7. Eui-Hong Han and Karypis G, Centroid-based Document Classification: Analysis and Experimental Results, 2000.
8. Deerwester S, Dumais S T, Furnas G W, Landaurer T K and Harshman R. Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41(1):391-407, 1990.
9. Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, pages 177-196, 2001.
10. Jollife I T, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
11. C Rorres and H Anton. *Elementary Linear Algebra*, 9^{th} Edition Application Version.

Sumanta Guha is currently the Associate Professor, Asian Institute of Technology, Pathumthani, Thailand. His qualifications are BSc, MSc, University of Calcutta, Calcutta; Ph.D, Indian Statistical Institute, Calcutta; MS, Ph.D, University of Michigan, Ann Arbor.