ICA and SOM in Text Document Analysis

Ella Bingham, Jukka Kuusisto and Krista Lagus
Neural Networks Research Centre, Helsinki University of Technology, Finland
ella@iki.fi, jukka.kuusisto@hut.fi, krista.lagus@hut.fi

ABSTRACT

In this study we show experimental results on using Independent Component Analysis (ICA) and the Self-Organizing Map (SOM) in document analysis. Our documents are segments of spoken dialogues carried out over the telephone in a customer service, transcribed into text. The task is to analyze the topics of the discussions, and to group the discussions into meaningful subsets. The quality of the grouping is studied by comparing to a manual topical classification of the documents.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*; I.5.4 [**Pattern Recognition**]: Applications—*Text processing*

General Terms

Algorithms, Experimentation

1. INTRODUCTION AND METHODS

The automatic analysis of textual documents and their topics is one of the challenges of modern information processing systems. One of the applications of this analysis is information retrieval. Given a set of documents, the task is to find out what a document is about, and which documents share similarities.

The methods used in this study are the Self-Organizing Map (SOM) and Independent component analysis (ICA). Both are completely unsupervised methods: when grouping a set of documents into subsets, no labeled training data is needed to aid in forming the groups.

ICA [2, 5] is a method for presenting a set of multivariate observations as a (linear) combination of unknown latent variables that are statistically independent. ICA was originally developed for signal processing purposes, but lately it has been found out that it is a suitable method for analyzing text documents, too, if documents are presented using the vector space model. The latent variables are in this case the document topics, and these can be regarded as probability distributions on the universe of terms. First approaches of finding the latent topics of a set of documents using ICA were presented in [6, 11].

It has been shown that the SOM [9] can be used to automatically organize very large document collections onto document maps [10]. On the ordered two-dimensional map, documents that are similar in content are found near each other. Using the SOM in text retrieval is described in [12].

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

2. DATA

The data used in the experiments were Finnish dialogues, recorded from the customer service phone line of Helsinki City Transport. The data was manually transcribed into text and base forms of the words were found automatically — in a morphologically rich language such as Finnish, this task is not straightforward. The data was provided by the Interact project [7]. In total, 57 dialogues were collected. A number of topic categories were selected so that they comprehensively encompass the subjects occurring in the data. The dialogues were manually tagged and segmented, so that each segment belongs to a topic category and forms a separate document. The total number of such documents was 195. The topic categories are 'timetables' (45 documents), 'beginnings' (57), 'tickets' (18), 'endings' (55) and 'out of domain' (20).

Characteristic of the transcribed data is that it is extremely colloquial: Both the customer and the customer servant use a lot of expletive words, such as 'nii' ('so', 'yea') and 'tota' ('hum', 'er', 'like'), often the words appear in reduced or otherwise non-standard forms, the word order does not always follow grammatical rules and quite frequently there is considerable overlap between the dialogue turns.

3. EXPERIMENTS AND DISCUSSION

The documents were encoded as vectors using the methods described in [10]. In short, the encoding was as follows. Stopwords (function words etc.) and words that appeared fewer than 2 times were removed. The remaining 1894 words were weighted using their entropy over manual document classes (cf. [10]). The documents were then encoded using the vector space model by Salton [14].

Furthermore, sparse random projection of the word vectors was applied to reduce the dimensionality of the vectors [10, 8]: for each word, 5 random dimensions out of a 500-dimensional vector were set to one, the rest being zeros. A document vector was then calculated as the sum of the vectors of its words, weighted by the word weights.

Results on ICA. We considered both the original 1894-dimensional data and 500-dimensional random projected data, as described above. The use of random projection as a means of speeding up computation in the context of ICA of text documents has not been reported elsewhere, up to our knowledge.

As a preprocessing step, the dimensionality of the term by document matrix was reduced using Latent Semantic Indexing [3] from the original high dimensionality (1894 or 500) to 10 or 15, which were the numbers of estimated topics.

Table 1: **Classification accuracies and CPU times of ICA and SOM.**

Method	Classification accuracy of topic categories, %					Total	CPU time, seconds
	timetables	beginnings	tickets	endings	out of domain	accuracy, %	(ICA in Matlab, SOM in C)
ICA 10	93.3	100	94.4	98.2	0	87.2	8.04
ICA RP 10	84.4	99.5	94.4	96.4	0	84.5	7.12
ICA 15	87.0	100	62.2	97.4	48.00	87.4	18.37
ICA RP 15	67.2	99.7	81.9	93.2	53.0	83.9	11.41
SOM RP	93.3	100	77.8	100	60.0	92.3	7.10

The FastICA algorithm[1] [4] was used to estimate the latent topics. As an output of the algorithm we get projection directions into which the data must be projected to reveal the latent variables (topics). The projections now tell the topic activities in documents; i.e. how good an example an observed document is of the topic.

Each estimated latent topic variable is mapped to the manual document category whose documents have the highest sum of topic activities in the latent variable. Also, each document is classified to that latent variable in which the document has highest topic activity. If the document is classified to a latent variable representing a different category than where the document was manually grouped, we consider the document misclassified. The total classification accuracy is the percentage of correct classifications.

The numbers of estimated topics (10 and 15) were found by trial and error. A smaller number than 10 lead to smaller classification accuracy; but as we see in Table 1, estimating 15 topics did not increase the accuracy but only required more computing power.

Results on the SOM. In our experiment, a SOM of $6 \times 4 = 24$ units was organized. Each document was placed into its best-matching map unit. Next, each map unit was assigned a topic by a majority vote of the documents in that unit. When the minority documents were counted as errors, the accuracy was 92.3% on the whole data set.

A size of 6×4 was chosen in order to have a non-square map grid where the proportion between number of documents and number of map units is roughly 1/10.

Comparison of results and discussion. A summary of the results and CPU times is presented in Table 1. The ICA results are averages over 20 runs, as different initializations may cause fluctuations. 10 or 15 latent ICA topics were estimated, and the data was either used in its original or random projected (RP) form; SOM only used random-projected data. The CPU times are not fully comparable as SOM uses C code and FastICA is realized in Matlab, and the implementations are somewhat different.

Both methods successfully classified large topic categories, especially 'beginnings' and 'endings'. In these categories, the documents are quite similar within the category, which may help the classification. The category 'out of domain' is hard to classify especially on ICA as there are only a few documents, and they do not form a statistically meaningful and coherent entity.

The estimated topic groups can be analyzed by their keywords. In the contexts of ICA and SOM, these are found as described in [1] and [13], respectively. The keywords (not listed here due to space limitations) nicely correspond to the true topic categories.

As a conclusion, ICA and SOM perform quite similarly in document analysis of a small dialogue data. A larger corpus with more categories would perhaps be needed to distinguish between the methods.

[1] `http://www.cis.hut.fi/projects/ica/fastica/`

4. REFERENCES

[1] E. Bingham. Topic identification in dynamical text by extracting minimum complexity time components. In *Proc. ICA2001*, pages 546–551, 2001.

[2] P. Comon. Independent component analysis — a new concept? *Signal Processing*, 36:287–314, 1994.

[3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[4] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Tr. on Neural Networks*, 10(3):626–634, May 1999.

[5] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis.* Wiley Interscience, 2001.

[6] C. L. Isbell and P. Viola. Restucturing sparse high dimensional data for effective retrieval. In *Adv. in Neural Inf. Proc. Systems 11*, pages 480–486, 1998.

[7] K. Jokinen. Sigdial — the USIX Interact project: Adaptivity in dialogue systems. *Elsnews*, 10(2):10, Summer 2001.

[8] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. IJCNN'98*, volume 1, pages 413–418. 1998.

[9] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences.* Springer, Berlin, 1995.

[10] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela. Organization of a massive document collection. *IEEE Tr. on Neural Networks*, 11(3):574–585, May 2000.

[11] T. Kolenda, L. K. Hansen, and S. Sigurdsson. Independent components in text. In M. Girolami, editor, *Advances in Independent Component Analysis*, chapter 13, pages 235–256. Springer-Verlag, 2000.

[12] K. Lagus. Text retrieval using self-organized document maps. *Neural Processing Letters*, 15(1):21–29, 2002.

[13] K. Lagus and S. Kaski. Keyword selection method for characterizing text document maps. In *Proc. ICANN99*, volume 1, pages 371–376, 1999.

[14] G. Salton and M. McGill. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983.