

# A SOM variant based on the Wilcoxon test for document organization and retrieval

A. Georgakis, C. Kotropoulos, and I. Pitas

Department of Informatics,  
Aristotle University of Thessaloniki,  
Thessaloniki 54006, Greece.  
e-mail: {apostolos, costas}@zeus.csd.auth.gr

**Abstract.** A variant of the self-organizing maps algorithm is proposed in this paper for document organization and retrieval. Bigrams are used to encode the available documents and signed ranks are assigned to these bigrams according to their frequencies. A novel metric which is based on the Wilcoxon signed-rank test exploits these ranks in assessing the contextual similarity between documents. This metric replaces the Euclidean distance employed by the self-organizing maps algorithm in identifying the winner neuron. Experiments performed using both algorithms demonstrates a superior performance of the proposed variant against the self-organizing map algorithm regarding the average recall-precision curves.

## 1 Introduction

Document organization and retrieval has been a vivid research and development area for the past 30 years with goals spanning from: *indexing* and *retrieval* to *representation* and *categorization* [1]. A fundamental problem in the area is the evaluation of the contextual similarity between documents. This paper describes a method for evaluating the contextual similarity between documents by addressing the issue of a proper distance between texts. In doing so we assume that the contextual similarity between documents exists also in their vectorial representation. Subsequently, the above mentioned similarity can be assessed through the use of a vector norm. For this purpose, the available textual data are represented by vectors using the *vector space* model [1, 2].

A plethora of document organization and retrieval systems are based on the vector space model. One system capable of clustering documents according to their contextual similarity is the well-known *Self-Organizing Map* (SOM) or *Kohonen* algorithm [3, 4]. In this paper, a variant of the SOM algorithm will be presented which is based on a novel vector norm.

A modeling technique based on the vector space model is used in order to effectively encode the documents. Subsequently, a new metric based on the aforementioned modeling technique and the Wilcoxon *signed-rank* test is introduced in order to assess the above mentioned similarity. Finally, the modeling method and the norm proposed are used in constructing a document organization system.

In what follows, Section 2 provides a brief description of the language modeling method. Section 3 provides a detailed description of the proposed norm, whereas, the next Section contains a brief description of the SOM algorithm. Section 5 describes the variant under consideration. Finally, in Section 6 we assess the effectiveness of both algorithms by using document-based queries.

## 2 Vector Construction

Let us suppose that we have a training corpus. The documents of the corpus are encoded into numerical vectors using the well-know *bigram* model [2]. For this purpose the maximum likelihood estimates of the conditional probabilities for the bigrams are computed as follows:  $x_{lm} = n_{lm}/N_l, \forall l, m \in \{1, 2, \dots, N\}$ , where  $n_{lm}$  is the number of times the bigram ( $l$ th word stem,  $m$ th word stem) occurred in the corpus,  $N_l$  is the number of times the  $l$ th word stem occurred in the corpus and  $N$  is the number of word stems in the corpus [2]. The *feature vector* corresponding to the  $i$ th document is given by:

$$\tilde{\mathbf{x}}_i = \sum_{l=1}^N \sum_{m=1}^N x_{lm} \mathbf{e}_{lm}, \quad (1)$$

where  $\mathbf{e}_{lm}$  denotes the  $(N^2 \times 1)$  unit norm vector having one in the  $(l \times N + m)$ th position and zero elsewhere. Let  $\mathbf{b}_i$  denote the *indicator vector* that contains the bigrams of the  $i$ th document.

To reduce the dimensionality in both vectorial types, the elements of the feature vectors are sorted into descending order and the same permutations are performed on the elements of the indicator vectors. Afterwards, a threshold, which is denoted by  $n$ , is used to divide both vectors into two parts. The first part in both vectors contains the most significant elements and is preserved while the second part contains the non-significant elements of the vectors and is rejected.

## 3 Document Distance

The proposed modeling method is based on the following assumption: two documents, that are contextually similar, with high probability, contain the same set of bigrams. To assess the degree of similarity between documents a new metric is introduced which is based on the distance between the entities of the same bigrams inside the indicator vectors corresponding to the documents.

Let us denote by  $\mathbf{b}_i$  and  $\mathbf{b}_k$  the indicator vectors corresponding to the  $i$ th and  $k$ th documents, respectively. The distance between two entities of the same bigram in the indicator vectors is given by:

$$d_{ik}(j) = \begin{cases} j - l, & \text{if } b_{i(j)} = b_{k(l)} \\ n^2, & \text{else,} \end{cases} \quad (2)$$

where  $b_{i(j)}$  corresponds to the  $j$ th bigram (in the order of their frequencies) in the  $i$ th document.

Subsequently, the distances obtained for all the bigram pairs are transformed into their absolute values, that is,  $d_{ik}^*(j) = |d_{ik}(j)|, \forall j \in \{1, 2, \dots, n\}$ . The absolute values are ranked into descending order, with tied ranks included where appropriate. Let  $r_{ik}^*(j)$  denote the rank corresponding to the absolute value  $d_{ik}^*(j)$  and  $n_z(ik)$  denote the number of zeros encountered when evaluating Eq. (2) for the  $i$ th and  $j$ th documents respectively. No ranks are assigned to the zeros resulted from Eq. (2). Let also  $N_z(ik)$  denote the total number of non-zero unsigned ranks (*i.e.*,  $N_z(ik) = n - n_z(ik)$ ).

The last step is the computation of the signed ranks, defined by:

$$r_{ik}(j) = \begin{cases} r_{ik}^*(j), & \text{if } d_{ik}^*(j) = d_{ik}(j) \\ -r_{ik}^*(j), & \text{if } d_{ik}^*(j) = -d_{ik}(j) \\ 0, & \text{if } d_{ik}(j) = 0. \end{cases} \quad (3)$$

Let us denote by  $W_{ik}^+$  and  $W_{ik}^-$  the sum of the positive and negative signed ranks, respectively. The distance between two documents is defined as  $W_{ik} = \min\{W_{ik}^-, W_{ik}^+\}$  and will be called the *Wilcoxon* distance, henceforth.

The Wilcoxon distance is the proposed metric and is employed in a hypothesis testing to assess whether the  $i$ th and the  $j$ th documents are contextually similar or not.

The null hypothesis is true (the documents are similar) when both indicator vectors consist of the same set of bigrams. In that case, due to the association between the bigrams and their frequencies, the identical bigrams are expected to be located at the same positions in the indicator vectors and the value of  $W_{ik}$  is expected to be near zero. Subsequently, the null hypothesis is accepted if the absolute value of  $W_{ik}$  is bounded by the *critical values*. In that case, the documents are contextually *relevant*, otherwise, they are *irrelevant*.

For  $N_z(ik) < 25$  the critical values of the Wilcoxon test can be found in any statistical book, whereas, when the number of non-zero unsigned ranks exceeds 25 the Wilcoxon test is approximated by the normal distribution. The parameters of the distribution are:  $\mu(ik) = (N_z(ik) \times (N_z(ik) + 1)) / 4$  and  $\sigma^2(ik) = (N_z(ik) \times (N_z(ik) + 1) \times (2N_z(ik) + 1)) / 24$ . In that case the critical values are derived from the table of the cumulative distribution function of the normal distribution.

## 4 Self-Organizing Maps

The SOMs are feedforward, artificial neural networks (ANN). Each neuron is equipped with a *reference* vector which is updated every time a new *feature* vector is assigned to that particular neuron. Let  $\mathcal{W}$  denote the set of reference vectors  $\{\mathbf{w}_l(p) \in \mathbb{R}^n, l = 1, 2, \dots, Q\}$ , where the parameter  $p$  denotes discrete time and the notion  $Q$  corresponds to the total number of neurons. During the training phase, the algorithm tries to identify the *winning* reference vector  $\mathbf{w}_s(p)$  to a specific feature vector  $\tilde{\mathbf{x}}_h$ . The index of the winning reference vector is given by  $s = \min \|\tilde{\mathbf{x}}_h - \mathbf{w}_l(p)\|$ , where  $\|\cdot\|$  denotes the Euclidean distance.

The reference vector of the winner as well as the reference vectors of the neurons in its neighborhood are modified towards  $\tilde{\mathbf{x}}_h$  using the following equation

$\mathbf{w}_i(p+1) = \mathbf{w}_i(p) + a(p) \times [\tilde{\mathbf{x}}_h - \mathbf{w}_i(p)]$ , where  $a(p)$  corresponds to the *learning rate* which is a monotonically decreasing parameter.

## 5 Wilcoxon Variant

In the proposed variant we replace the Euclidean norm with the Wilcoxon distance in identifying the winner neuron. Let  $S_q(p)$  denote the set of indicator vectors that have been assigned to the  $q$ th neuron until the  $p$ th iteration. Let also  $\mathbf{b}_{q_{vm\text{ed}}} \in S_q(p)$  denote the so-called *vector median* corresponding to the set  $S_q(p)$  [5]. The vector median corresponds to the indicator vector that minimizes the  $L_1$  norm over the set  $S_q(p)$ :

$$\sum_{i=1}^{|S_q(p)|} W_{iq_{vm\text{ed}}} \leq \sum_{i=1}^{|S_q(p)|} W_{ij} \quad \forall j = 1, 2, \dots, |S_q(p)|, \quad (4)$$

where  $|S_q(p)|$  denotes the *cardinality* of the set  $S_q(p)$ . The vector median corresponding to the set  $S_q(p)$  stands for the reference vector of the  $q$ th neuron ( $\mathbf{w}_q(p) \equiv \mathbf{b}_{q_{vm\text{ed}}}$ ).

In identifying the index of the winner neuron with respect to a specific, randomly selected, indicator vector  $\mathbf{b}_h$ , the Wilcoxon distances between all the reference vectors of the ANN and the indicator vector under consideration are assessed. If the null hypothesis related to the Wilcoxon distance is validated to be “true”, then the reference vector corresponds to the winner neuron. It must be noted that it is possible multiple winners to stem out from the above procedure.

Let  $s$  denote the index of a winner neuron that stemmed from the above procedure. The set  $S_s(p)$  is updated with the vector  $\mathbf{b}_h$ , that is,  $S_s(p) = S_s(p) \cup \mathbf{b}_h$  and the corresponding vector median,  $\mathbf{b}_{s_{vm\text{ed}}}$ , is also updated. At the completion of each iteration, that is, when all the indicator vectors have been presented to the network, the sets of reference vectors are updated as follows:  $S_q(p+1) = \{\mathbf{w}_s(p)\}, \forall q = 1, 2, \dots, Q$ . Finally, prior to the completion of the training phase and in order to fine tune the map, the identification of the winning neuron is achieved by:  $s = \min \{W_{hq}\}, \forall q = 1, 2, \dots, Q$ . The above equation results in assigning the bigram vector  $\mathbf{b}_h$  to only one neuron during each iteration.

## 6 Document Organization and Retrieval

To test the proposed Wilcoxon variant against the SOM algorithm the Reuters-21578 corpus was used, which is an annotated corpus [6]. The SGML tags, the URLs, the email addresses, and the punctuation marks were removed. Subsequently, some common words and frequent terms were removed also and *stemming* was performed. Finally, the documents were encoded into numerical vectors.

These vectors are presented iteratively an adequate number of times to each one of the NNs in an effort to construct clusters containing semantically related documents. This process yields the so-called *document map* (DM) [3]. The DM

**Fig. 1.** The document map constructed for the Reuters-21578 for a  $9 \times 9$  neural network using the Wilcoxon variant. The highlighted neurons correspond to document clusters related to “financial debts” (top middle and left), “bonds” (bottom left) and “corporate economic results” (bottom right).

corresponding to the Reuters-21578 corpus using the Wilcoxon variant can be seen in Fig. 1. Each hexagon on the DM corresponds to one document category and the levels of grey correspond to different document densities. Hexagons with grey levels near 255 imply that fewer documents have been assigned to these neurons, whereas, grey levels near 0 imply higher document densities.

To evaluate the performance of the algorithms, with respect to their document organization capabilities, document-queries are used. For each query, the algorithms identify the winning neurons on the computed DMs and retrieve the documents of the training corpus associated with the winners. Subsequently, the retrieved documents are ranked according to their distance from the queries using either the Euclidean or the Wilcoxon distance. Finally, the retrieved documents are labeled as either relevant or not to the document-query, with respect to the annotation category they bear.

The relevance between the retrieved documents and the queries leads to the partitioning of the training corpus into two sets, one containing the relevant documents and another with the non-relevant documents. The effectiveness of the algorithms is assessed using the *average recall-precision* curve [7]. Figure 2a and Fig. 2b depict the *eleven-point* average recall-precision curves for the standard SOM and the Wilcoxon variant for the topics with the highest frequencies. At the beginning the performance of the SOM algorithm was slightly better

(a)

(b)

**Fig. 2.** (a) The recall-precision curve for the standard SOM and the Wilcoxon variant for the “Mergers and Acquisitions (ACQ)” category. (b) The recall-precision curves for both algorithms for the “Earnings and Earnings Forecasts (EARN)” category.

than the proposed variant but it degrades rapidly as the volume of the retrieved document grows.

## 7 Conclusions

A variant of the SOM algorithm for document organization and retrieval has been presented in this paper. The Euclidean distance used by the SOM algorithm in identifying the winning neuron is replaced by a novel metric which exploits the correlation between the words formulating the documents. The performance of the proposed variant with respect to the average recall-precision curves have been demonstrated to be superior than the SOM algorithm. Further investigations will be made towards the enhancement of the suggest algorithm in exploiting the latent textual information.

## References

1. R. B. Yates and B. R. Neto, *Modern Information Retrieval*, ACM Press, 1999.
2. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999.
3. T. Kohonen, *Self Organizing Maps*, Germany: Springer-Verlag, 1997.
4. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, V. Paatero, and A. Saarela, “Organization of a massive document collection,” *IEEE Trans. on Neural Networks*, vol. 11, no. 3, pp. 574–585, May 2000.
5. J. Astola, P. Haavisto, and Y. Neuro, “Vector median filters,” *Proceedings of the IEEE*, vol. 78, no. 4, pp. 678–689, April 1990.
6. D. D. Lewis, “Reuters-21578 text categorization test collection, distribution 1.0,” Sep. 1997, <http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>.
7. R. R. Korfhage, *Information Storage and Retrieval*, New York: J. Wiley, 1997.