A combination of Wilcoxon test and R-estimates for document organization and retrieval

A. Georgakis, C. Kotropoulos, and I. Pitas
Department of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
e-mail: {apostolos, costas, pitas}@zeus.csd.auth.gr

ABSTRACT

The Wilcoxon signed-rank test is exploited for document organization and retrieval in this paper. A novel modeling method for documents and a distance metric between documents are proposed. Both document modeling and document comparisons are based on signed-ranks and are applied to the frequency of occurrence of the document bigrams. A metric using the Wilcoxon signed-rank test exploits these signed-ranks to assess the contextual similarity between the documents. A variant of the self-organizing map algorithm (SOM), the so-called Wilcoxon SOM, is built. The quality of the document clustering produced by the Wilcoxon SOM against the standard SOM is compared in terms of their average recall-precision rates for document retrieval based on the document clusters formed. The reported improvements in recall and precision are shown to be statistically significant.

1 Introduction

A crucial issue for the efficient document organization and retrieval is the evaluation of the contextual similarity between documents. This paper provides a method for evaluating the document similarity by addressing the issue of a proper distance between documents represented as vectors in a vector space model [1]. Furthermore, it provides a novel modeling technique based on the ranking of the fundamental elements of a document, namely its bigrams (*i.e.*, pairs of consecutive words), according to their frequency of occurrence [1].

Given a document similarity measure and a clustering algorithm, one can easily partition a collection of documents, called *corpus*, into clusters of contextually relevant documents. The *Self-Organizing Map* (SOM) algorithm will be used in this paper [2, 3]. SOMs are neural networks (NNs) that employ a layer of input neurons and a single computational layer. The neurons on the computational layer are arranged on a N-dimensional lattice. In this paper, a variant of the SOM algorithm shall be used. This variant exploits the Wilcoxon signed-rank test to cluster contextually similar documents into classes.

The outline of the paper is as follows. Section 2 provides a brief description of the language modeling scheme. Section 3 provides a detailed description of the Wilcoxon metric. The SOM algorithm is presented in Section 4 and the Wilcoxon SOM variant in Section 5. Section 6 assess the effectiveness of both algorithms by using document-based queries and average recall-precision curves. Finally, in Section 7 we demonstrate that the improvements in recall-precision rates reported are statistically significant.

2 Language modeling

The documents of the corpus are encoded into numerical vectors using the well-know *bigram* model [1]. Let $V = \{w_1, w_2, \ldots, w_N\}$ denote the set of word types found in the corpus, where N denotes the vocabulary size. Let also $x_{lm} = P(w_m|w_l)$, $\{l, m = 1, 2, \cdots, N\}$ denote the conditional probability of the bigram (w_l, w_m). We construct the so-called *feature* vector that holds the conditional probabilities of the bigrams found in the ith document of the training corpus:

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

where $\mathbf{e}_{lm}$ denotes the $(N^2 \times 1)$ unit vector having one in the $(l \times N + m)$th entry and zero elsewhere.

Let also $\tilde{\mathbf{b}}_i$ denote an *indicator* vector of the form:

$$\tilde{\mathbf{b}}_i = (\delta_{11}b_{11}, \cdots, \delta_{1N}b_{1N}, \cdots, \delta_{NN}b_{NN})^T \qquad (2)$$

that contains the bigrams found in the ith document and zero otherwise.

Subsequently, the elements of $\tilde{\mathbf{x}}_i$ are arranged in descending order of the conditional probabilities. Similarly, the elements of $\tilde{\mathbf{b}}_i$ are permuted so that they appear with the same order of appearance as in the feature vector $\tilde{\mathbf{x}}_i$. The resulting vector is defined as:

$$\tilde{\mathbf{b}}_i = \left(b_{i(1)}, b_{i(2)}, \cdots, b_{i(N^2)}\right)^T \qquad (3)$$

where $(\cdot)$ indicates order statistics, $i.e.$, $b_{i(1)}$ indicates the most frequently occurred bigram in the ith document and so on. To reduce the dimensionality of both vectors, we may retain the n most frequent bigrams and confine the analysis to $(n \times 1)$ feature and indicator vectors, henceforth. Let us denote by $\mathbf{x}_i$ and $\mathbf{b}_i$ the preserved vector parts of the feature vector and the indicator vector respectively, after sorting and thresholding.

3 Wilcoxon distance

The proposed document similarity measure 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 based on the distance between the entities of the same bigrams inside the indicator vectors is introduced.

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

$$d_{ik}(j) = \begin{cases} j - l, & \text{if } b_{i(j)} = b_{k(l)} \\ n^2, & \text{else,} \end{cases} \qquad (4)$$

where $b_{i(j)}$ corresponds to the jth bigram in the ith document.

Subsequently, the distances obtained for all the bigram pairs are transformed into their absolute values. 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 of $d_{ik}(j)$ and $n_z(ik)$ denote the number of zeros encountered when evaluating Eq. (4) for the ith and kth documents. No ranks are assigned to the zeros that result from Eq. (4). Let also $\zeta(ik)$ denote the total number of non-zero unsigned ranks ($i.e.$, $\zeta(ik) = n - n_z(ik)$).

The signed ranks are computed by:

$$\begin{array}{ll} r_{ik}^+(j) = r_{ik}(j), & \text{if } d_{ik}(j) > 0 \\ r_{ik}^-(j) = r_{ik}(j), & \text{if } d_{ik}(j) < 0. \end{array} \qquad (5)$$

Let $W_{ik}^+ = \sum_{j=1}^n r_{ik}^+(j)$ and $W_{ik}^- = \sum_{j=1}^n r_{ik}^-(j)$ denote the sum of the positive and negative signed ranks, respectively. The distance between the ith and kth document 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 whether the ith and the kth document are contextually similar or not. The null hypothesis is:

- H_0: The two documents have the same contextual and semantical content.

The null hypothesis is true when both indicator vectors consist of the same set of bigrams. In this case, due to the association between the bigrams and their frequencies, identical bigrams are expected to be located at the same positions or close together in the indicator vectors and the value of W_{ik} is expected to be near zero. Subsequently, the null hypothesis is accepted if W_{ik} is bounded above by the $critical\ values$ of the Wilcoxon test. In this case, the documents are marked as contextually $relevant$, otherwise, they are labeled as $irrelevant$.

For $\zeta(ik) < 25$ the critical values of the Wilcoxon test can be found in any nonparametric 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: $\mu(ik) = \zeta(ik)(\zeta(ik) + 1)/4$ and $\sigma^2(ik) = \zeta(ik)(\zeta(ik) + 1)(2\zeta(ik) + 1)/24$. In this case, the critical values are derived from the table of the cumulative distribution function of the normal distribution.

4 Self-Organizing Maps

Let $\mathcal{W}$ denote the set of reference vectors of the neurons on the computational layer of the SOM, $\{\mathbf{w}_l(t) \in I\!R^n, l = 1, 2, \ldots, Q\}$, where the parameter t denotes discrete time (number of iterations) and Q corresponds to the total number of neurons. During the training phase, the algorithm tries to identify the $winning$ reference vector, $\mathbf{w}_s(t)$, to a specific feature vector $\mathbf{x}_h$. The index of the winning reference vector is given by: $s = \arg\min \|\mathbf{x}_h - \mathbf{w}_l(t)\|$, $l = 1, 2, \ldots, Q$, 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 $\mathbf{x}_h$ using the following equation: $\mathbf{w}_i(t+1) = \mathbf{w}_i(t) + a(t)\left[\mathbf{x}_h - \mathbf{w}_i(t)\right]$, where $a(t)$ is the $learning\ rate$, which is a monotonically decreasing parameter and the index i denotes the neurons in the neighborhood of the winner neuron.

5 Wilcoxon Variant

Let the weight vector of each neuron, during the initialization phase of the algorithm, be chosen as one of the available indicator vectors. In the Wilcoxon SOM variant the Euclidean norm is replaced with the Wilcoxon distance in identifying the winner neuron.

Let $S_q(t)$ denote the set of indicator vectors that have been assigned to the qth neuron until the tth time instant. Let also $\mathbf{b}_{q_{vmed}} \in S_q(t)$ denote the Wilcoxon *vector median* (VM) corresponding to the set $S_q(t)$. The Wilcoxon VM is a variant of the vector median proposed in [4]. The vector median corresponds to the indicator vector that minimizes the sum of Wilcoxon distances between each vector in $S_q(t)$ and any other vector in $S_q(t)$:

$$\sum_{i=1}^{|S_q(t)|} W_{iq_{vmed}} \leq \sum_{i=1}^{|S_q(t)|} W_{ij}, j = 1, 2, \ldots, |S_q(t)|, \quad (6)$$

where $|\cdot|$ denotes the *cardinality* of the set $S_q(t)$. The Wilcoxon VM stands for the weight vector of the qth neuron, *i.e.* $(\mathbf{w}_q(t) \equiv \mathbf{b}_{q_{vmed}})$.

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 NN 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 one or more than one winner neurons can be identified from such a test.

After the identification of the winner neuron and for either single or multiple winners the updating procedure is carried out as follows: the corresponding set of indicator vectors $S_s(t)$ is updated with the vector $\mathbf{b}_i$, that is, $S_s(t) = S_s(t) \bigcup \{\mathbf{b}_i\}$, and the vector median corresponding to the sth neuron is also updated using Eq. (6).

With the advance of time the significance level in the hypothesis testing for the winner identification is increased from 0.01 to 0.025 and finally to 0.05. As a result, fewer neurons are labeled as winner neurons. Multiple assignment of the same document, during the early iterations of the algorithm, are permitted in order to achieve a global ordering of the indicator vectors and accordingly the documents on the lattice.

The weight vector of each neuron in the origin of the $(t+1)$th iteration is set to be equal to the vector median found during the tth iteration ($\mathbf{w}_q(t+1) \equiv \mathbf{b}_{q_{vmed}}(t)$, $q = 1, 2, \ldots, Q$) and the sets S_q are also initiated as follows:

$$S_q(t+1) = \{\mathbf{w}_q(t)\}, q = 1, 2, \ldots, Q. \quad (7)$$

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:

$$s = \min\{W_{iq}\}, \forall q \in \{1, 2, \ldots, Q\}, \quad (8)$$

that is, we select the neuron whose Wilcoxon distance from the indicator vector under consideration is minimum. Eq. (8) can be used as an ultimate mean to resolve ambiguities when multiple winners are still determined.

6 Document Organization and Retrieval

To test the proposed Wilcoxon metric against the SOM algorithm the Reuters-21578 corpus was used [5]. The SGML tags 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 vectors using both vectorial types, that is, the feature vectors and the indicator vectors.

These vectors are presented iteratively an adequate number of times to each one of the NNs build using either the Euclidean or the Wilcoxon metric in an effort to construct clusters containing semantically related documents. This process yields the so-called *document map* (DM) [2]. The DM 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. The documents used for queries are drawn from the test set of the Reuters-21578 corpus according to the recommended *Modified Apte* split of the collection [5]. Each document-query passes the same preprocessing steps like the training documents and is encoded by a feature vector and an indicator vector. Following, the algorithms identify the winning neurons on the computed DMs and retrieve the set of 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.
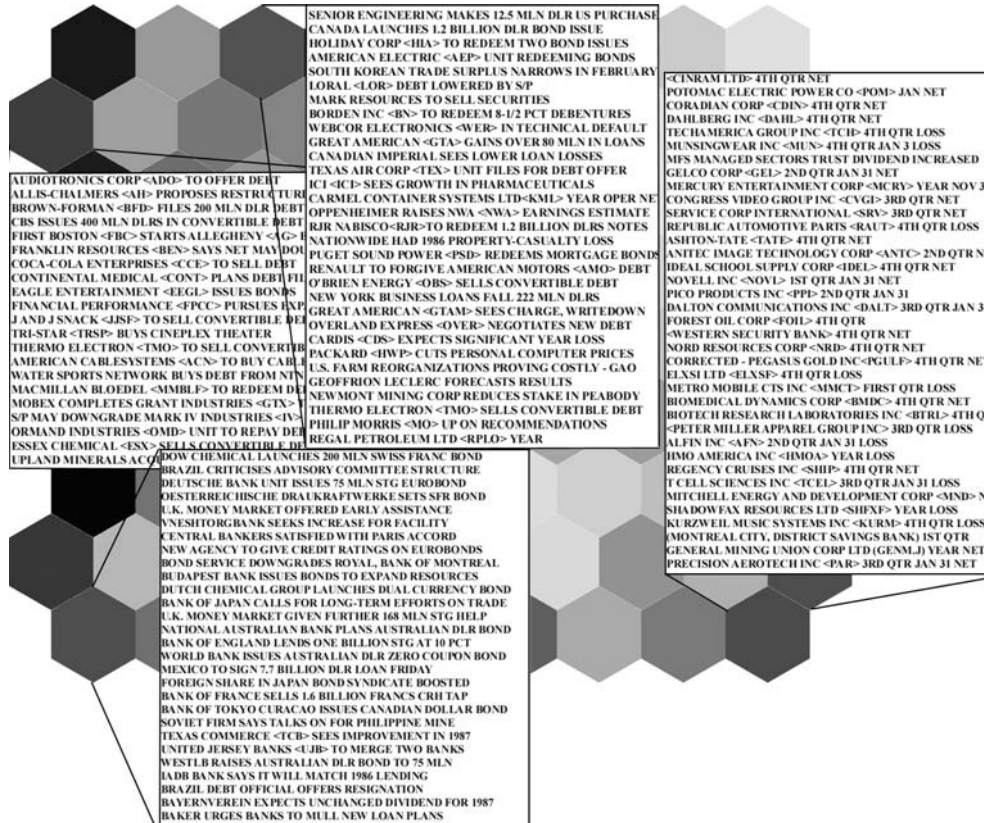
Figure 1: The document map constructed for the Reuters-21578 for a 9×9 neural network using the Wilcoxon variant. The highlighted neurons correspond to document clusters related to "financial debts" (top middle and left), "bonds" (bottom left), and "corporate economic results" (bottom right).

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 [6]. Figure 2a depicts the average recall-precision curves for the standard SOM algorithm and the Wilcoxon variant for the "Mergers and Acquisitions (ACQ)" category. The performance of the Wilcoxon variant is higher to the standard SOM algorithm in small recall volumes ($recall < 0.35$) by $2-5\%$, which is extremely important given the fact that an average user is interested in high precisions ratios even from the beginning of the list of returned relevant documents. Figure 2b depicts the average recall-precision curves for the standard SOM algorithm and the Wilcoxon variant for the "Earnings and Earnings Forecasts (EARN)" category. At the beginning the performance of the SOM algorithm is slightly better than the proposed

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

7 Statistical evaluation

To assess whether the observed differences in the evaluation between the proposed Wilcoxon SOM variant and the standard SOM algorithm are really meaningful or merely a chance, a hypothesis test is employed. Several hypothesis tests can be employed ranging from the t-test and the signed rank test to the Wilcoxon test and the one-way ANOVA test between two different methods [7].

Let us denote by P_i^w the precision score of the Wilcoxon SOM variant for the ith annotation category and by P_i^s the precision scores of the standard SOM algorithm for the same annotation category. To compare the performance between these two algorithms the paired Wilcoxon signed rank test is used. Let $D_i = P_i^w - P_i^s$ denote the precision differences between the Wilcoxon variant and the standard SOM algorithm for the ith annotation category. The hidden model assumed here is $D_i = \theta + e_i$, where e_i denotes the error that is taken independent of θ
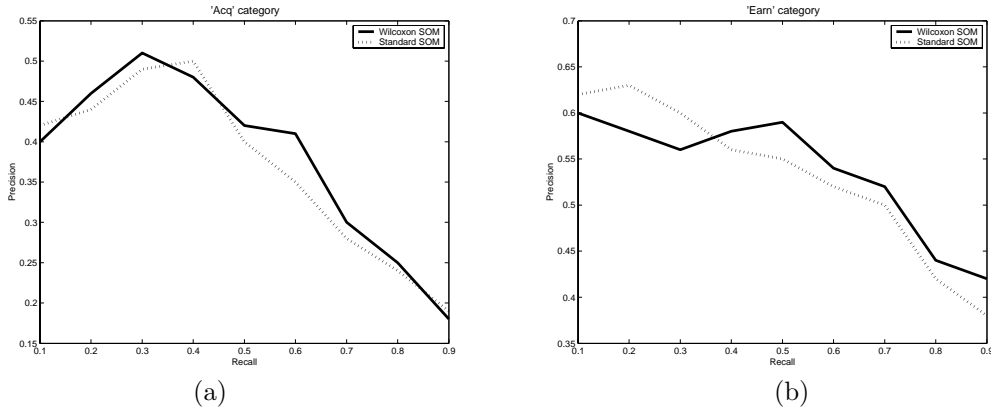
Figure 2: (a) The recall-precision curve for the standard SOM and the Wilcoxon variant for the "Mergers and Acquisitions (ACQ)" category. (b) The recall-precision curves for both algorithms for the "Earnings and Earnings Forecasts (EARN)" category.

with mean value equal to zero, that is, $E(e_i) = 0$. The null hypothesis to be tested is that $\theta = 0$ under the alternative hypothesis $\theta > 0$. If the alternative hypothesis is accepted, then there is a statistically significant improvement on the precision of the Wilcoxon SOM variant.

Table 1 contains the results of the hypothesis test under the 0.01, 0.025 and 0.05 levels of significance for each of the aforementioned annotation categories. It is evident that for the "acq" category the improved precision performance of the Wilcoxon SOM variant at each level of significance is statistically significant against that of the SOM algorithm. In the "earn" annotation category, this is true for 0.025 and 0.05 levels of significance.

8   Conclusions

A novel modeling method and a metric based on this method were introduced in this paper for document organization and retrieval. The modeling method relied partially on the frequency of appearance of the bigrams forming the documents and on the ranking of these bigrams according to these frequencies. Signed ranks were assigned to the ordered bigrams. The proposed metric relied on the Wilcoxon signed-rank test and exploited these ranks in assessing the contextual similarity between documents. These two novel techniques were incorporated in the SOM al-

Table 1: Precision performance evaluation between the Wilcoxon SOM variant and the standard SOM.

Category	Reject H_0		
	0.01	0.025	0.05
acq	true	true	true
earn	false	true	true

gorithm in identifying the winner neuron and their performance was tested against the standard metric used by the algorithm, that is, the Euclidean distance. The performance of the proposed variant with respect to the average recall-precision curves have been demonstrated to be statistically superior than the standard SOM algorithm.

References

[1] D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, Cambridge, MA: MIT Press, 1999.

[2] T. Kohonen, *Self Organizing Maps*, Berlin: Springer-Verlag, 1997.

[3] 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.

[4] J. Astola, P. Haavisto, and Y. Neuro, "Vector median filters," *Proceedings of the IEEE*, vol. 78, no. 4, pp. 678–689, April 1990.

[5] D. D. Lewis, "Reuters-21578 text categorization test collection, distribution 1.0," 1997, http://kdd.ics.uci.edu/databases/reuters21578/-reuters21578.html.

[6] R. R. Korfhage, *Information Storage and Retrieval*, New York: J. Wiley, 1997.

[7] D. Hull, "Using statistical testing in the evaluation of retrieval performance," in *Proc. of the 16th ACM/SIGIR Int. Conference on Research and Development in Information Retrieval*, 1993, pp. 329–338.