In Proc. of ICANN99, Ninth Int. Conf. on Artificial Neural Networks
vol. 1, pp. 371--376. IEE, London, 1999.

Keyword selection method for characterizing text document maps

Krista Lagus and Samuel Kaski
Helsinki University of Technology, Neural Networks Research Centre,
P.O. Box 2200 FIN-02015 HUT, Finland.
E-mail: Krista.Lagus@hut.fi

Abstract

Characterization of subsets of data is a recurring problem in data mining. We propose a keyword selection method that can be used for obtaining characterizations of clusters of data whenever textual descriptions can be associated with the data. Several methods that cluster data sets or form projections of data provide an order or distance measure of the clusters. If such an ordering of the clusters exists or can be deduced, the method utilizes the order to improve the characterizations. The proposed method may be applied, for example, to characterizing graphical displays of collections of data ordered e.g. with the SOM algorithm. The method is validated using a collection of 10,000 scientific abstracts from the INSPEC database organized on a WEBSOM document map.

1 Introduction

Graphical displays of collections of data have in the recent years gained popularity in data mining. Methods such as the self-organizing map are used to cluster or organize large amounts of data onto a map, which can then be visualized and used in interpretation and exploration of the data collection. Subsequently it has become important to develop methods that aid interpretation and facilitate visually guided exploration of such maps or clusterings. If textual descriptions of the data are available or if the data itself is textual, the descriptors may be *keywords* that characterize the sub-collection of text found in a particular region. When written on the graphical display, these words also function as *landmarks*, i.e., navigational cues that help in maintaining a sense of location during exploration of the map.

Somewhat related problems and methods such as *keyword extraction* and *term weighting* have been studied in the field of information retrieval (IR). However, there the problem usually is that of selecting or weighting terms that describe *a single document*, and a typical goal is that of *effective retrieval* of relevant documents from a large collection. In contrast, in landmark selection the objective is to provide descriptors and visual cues for a human exploring the collection of data. Furthermore, instead of describing individual documents, the goal is to characterize clusters or map areas containing several similar documents, and to find optimal places to be labeled. The novel problem area has emerged due to the ability to automatically construct ordered, graphical displays of large collections of data.

In this article we introduce a keyword selection method that can be utilized for suggesting descriptive terms, or *labels* for groups of similar documents or clusters of textual data. The method is especially suitable for characterizing maps of data collections organized using the self-organizing map (SOM) algorithm. The method focuses on the distributions of words occurring within the document groups. The method is validated using a document map that organizes a collection of scientific abstracts from the INSPEC database, where human-assigned lists of descriptive terms are available to provide a basis for comparison. We have applied the proposed method for characterizing text document maps constructed using the WEBSOM method.

1.1 Self-organizing maps

The self-organizing map (SOM) algorithm (Kohonen, 1982; Kohonen, 1995) is a means of automatically arranging high-dimensional statistical data so that alike inputs are in general mapped close to each other. A self-organizing map consists of a regular grid of *processing units* with associated models that are capable of representing data. Often a two-dimensional map grid is used for the sake of easy visualization. Map units that lie near each other on the grid are called *neighbors*. After constructing a self-organizing map for a data set, neighboring map units represent similar kinds of data items whereas more distant map units represent different kinds of data.

1.2 Document maps

Large collections of text documents can be organized onto *document maps* using the WEBSOM method (Honkela et al., 1996; Kaski et al., 1998; Kohonen et al., 1999; Lagus et al., 1999). On such maps, nearby areas contain documents that are mutually similar in content. The WEBSOM method utilizes the SOM algorithm to organize the documents onto a 2-dimensional map display, which provides the basis for interactive exploration of the document collection. The method proposed in this paper has been applied for automatic labeling of WEBSOM document maps (see `http://websom.hut.fi/websom/` for some demonstrations).

2 Methods

2.1 Keywords for clusters

The goodness of a keyword, or of any *descriptor*, can be described intuitively as follows:

A good descriptor of a cluster characterizes some outstanding property of the cluster in relation to the rest of the collection.

In other words, for a word w to be a good keyword for a document cluster C, w should have the following two properties:

1. w should be prominent in C *compared to other words in* C

2. w should be prominent in C *compared to the occurrence of* w *in the whole collection.*

For the purpose of ranking keywords these two criteria can be combined into the general measure

$$G(w) = F^{clust}(w) \times F^{coll}(w), \qquad (1)$$

where the first term, F^{clust}, describes the word w in relation to other words within the cluster j to be described, whereas the second term, F^{coll}, relates the word to the whole collection. There are naturally a multitude of possible measures that follow this general form. We will now propose one such measure.

Let $f_j(w)$ denote the number of times word w occurs in cluster j, i.e. the *frequency* of word w in j. Then, let $F_j(w)$ denote the *relative frequency* of word w, defined as

$$F_j(w) = \frac{f_j(w)}{\sum_v f_j(v)}. \qquad (2)$$

Note that $0 < F_j(w) < 1$ and $\sum_w F_j(w) = 1$. The effect of this normalization is to disregard the sizes of the clusters, and instead to measure the relative importance of a word compared to the other words occurring in the cluster. The relative frequency $F_j(w)$ now seems a good candidate for F_j^{clust}.

Next, we would like F_j^{coll} to measure the relation of the frequency of w in cluster j to the "background frequency" that describes how typical the word is in other parts of the collection. A straightforward measure for this comparison is $F_j(w)/\sum_i F_i(w)$. Now we have both of the components necessary for our goodness measure:

$$F_j^{clust} = F_j(w), \text{ and} \qquad (3)$$

$$F_j^{coll} = F_j(w)/\textstyle\sum_i F_i(w), \qquad (4)$$

where i and j are cluster indexes. The component F_j^{clust} favors words that take up a large proportion of their cluster, whereas F_j^{coll} inhibits words that appear in large proportions in other clusters as well.

Now we are ready to define the *goodness* G^0 of a word w appearing in cluster j as

$$G^0(w,j) = F_j(w)\frac{F_j(w)}{\sum_i F_i(w)}. \qquad (5)$$

2.2 Keywords for map units

What has been said earlier about keyword selection for clusters applies also in the case of selecting keywords for individual SOM units. Furthermore, since the map units have an order, i.e., the neighbors of a map unit are more similar to it than the units farther away, it may be useful to utilize also this readily available order information in the definition of a goodness value.

Moreover, if the word w occurs often in map unit j, the word is probably relatively common also in some adjoining area of the map. However, the frequent appearance of w close by does not make it a bad keyword for map unit j. Thus we would like to exclude an area of the map immediately surrounding unit j in the calculation of the "inhibitory factor" $F_j^{coll}(w)$ and to reformulate G^0 into a new definition of goodness:

$$G^1(w,j) = F_j(w)\frac{F_j(w)}{F_j(w) + \sum_{i \notin A_1^j} F_i(w)}, \quad (6)$$
$$\text{where } i \in A_1^j \text{ if } d(j,i) < r_1,$$

and $d(j,i)$ is the distance on map grid between units i and j. There A_1^j (see Fig. 1) is a "neutral map zone" around the unit j that will not participate in any way in determining the goodness of word w as a keyword for unit j.

2.3 Keywords for map areas

Due to the neutral zone introduced in Sec. 2.2 the measure of goodness of a keyword w for unit j does not punish words in unit j merely for being good descriptors also for the neighboring units. However, if we are looking for good descriptors for *larger map areas*, as in finding labels for large portions of the graphical map display, we would like to reward a word in unit j if it is a good descriptor of the neighboring units as well.

The goodness value G^1 can be re-expressed in a way that explicitly rewards several map units within radius r_0 for forming a cluster in terms of word w:

$$G^2(w,j) = [\sum_{k \in A_0^j} F_k(w)]\frac{\sum_{k \in A_0^j} F_k(w)}{\sum_{i \notin A_1^j} F_i(w)}, \quad (7)$$
$$\text{where } k \in A_0^j \text{ if } d(j,k) < r_0, \text{ and}$$
$$i \in A_1^j \text{ if } r_0 < d(j,i) < r_1,$$

and $d(j,i)$ is again the distance on map grid between units i and j, r_1 the radius of the

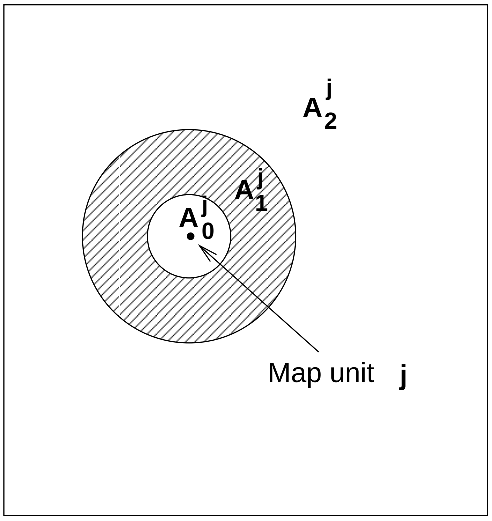

Figure 1: Determining the goodness value G^2 for words in map unit j. The shaded area (A_1^j) is disregarded when calculating the goodness values for each word in unit j.

"neutral zone", and r_0 the radius of the map area to be characterized by the keyword (see Fig. 1). It seems that G^2 implements our intuitions regarding measuring the goodness of a keyword as a descriptor for map area centered at map unit j. The computational load can be reduced by the following approximation: use G^1 to pre-select a fixed number of candidate keywords per map unit (saving some intermediate results), and then select final keywords based on calculating G^2.

2.4 Labels for graphical map displays

The labels on a graphical map display serve several different purposes. An individual label is a descriptor of the underlying area, guiding towards interesting information. Collectively the labels may serve as a summary of the data collection. Furthermore, if the map may be viewed using different resolutions (zooming), the labels function as *landmarks*: they help orienting by providing reference points during transitions across views having different resolutions.

Usually it is neither possible nor desirable to label every map unit on the display. Often there is not enough space on the graphical map display, and even if there were, cramming the display with masses of words should not be called visualization.

Therefore, to obtain the final labeling of a map, we need to place labels on a subset of the map units so that the resulting labeling is as good as possible. If the total goodness of the labeling is defined as sum of the goodnesses of all labels, there are $\binom{N}{M}$ possible combinations of N labels out of M candidates. Obviously all combinations cannot be evaluated in most practical situations. Fortunately, in this type of selection problems a greedy approach usually produces a near-optimal approximation.

With WEBSOM document maps that have several zooming levels, we have used the following procedure for selecting labels for each level l, starting with the topmost level. First, decide for each display level l the desired labeling density expressed in terms of minimum distance on map grid between two labeled units i and j, $d_l(i,j)$. Then, perform the following steps for each level:

1. Order map units according to G^2 of the best keyword in the unit.
2. Repeat: accept the best word from the best unit on the map if it is separated at least by distance d_l from all already accepted labels.
3. When no more labels can be added for level l, increment l.

We have obtained good results by choosing the radius parameters r_0 and r_1 used in calculating G^2 so that half of the desired labeling density, $d/2$, lies between r_0 and r_1.

3 Experiments

To validate the keyword selection method, we used a collection of $10,000$ scientific abstracts from the INSPEC database. In addition to the abstract, each document contained a list of keywords pre-assigned to the document. We used the abstracts for organizing a document map and for selecting keywords with the proposed method, and the lists of manually pre-assigned keywords to evaluate the performance of our method.

3.1 Construction of document map

Detailed descriptions of how to construct document maps can be found in (Kaski et al., 1998; Kohonen et al., 1999; Lagus et al., 1999) and therefore only some main choices are mentioned here. The abstracts (not including the pre-assigned keyword lists) were first encoded using Salton's vector space model (Salton et al., 1975). Words were weighted according to how well they differentiated classes (class-entropy based weighting). After encoding, the document vector dimension was reduced using random mapping of the vectors. The resulting vectors were then organized with the SOM algorithm onto a map of $1,040$ units (26×40).

3.2 Validation of proposed method

We wanted to compare the proposed keyword selection method against term lists independently assigned to the documents, to see how similar results our method would produce. As mentioned earlier, the abstracts were provided with term lists (the record fields called *"Free terms"* and *"Keywords"*), which were then used as the "ground truth" regarding keyword selection.

Since the term lists compiled by humans did not contain ordering of the terms by relevance, we utilized a standard term weighting method in forming ordered *validation list* of terms for each map unit. More specifically, all the terms provided with the abstracts in map unit j were collected into a list, and ordered according to the *validation value* $V_j(w)$:

$$V_j(w) = f_j(w) \times IDF(w), \tag{8}$$

where $f_j(w)$ is the number of occurrence of term w in unit j, and IDF is *inverse document frequency*, a classical term weighting approach used in information retrieval systems (Salton and Buckley, 1987; Church and Gale, 1995). It is defined as $\log \frac{N}{N_w}$, with N the number of documents in the collection, and N_w the number of documents containing the term w. Note that the validation lists were still manually pre-assigned: IDF was used only in ordering the lists.

Next, we compared for each map unit the best keywords selected by our method with the best-ranking terms in the validation list of the unit. Let the best word produced by our method be called b_j and the validation list with N highest-ranked keywords from

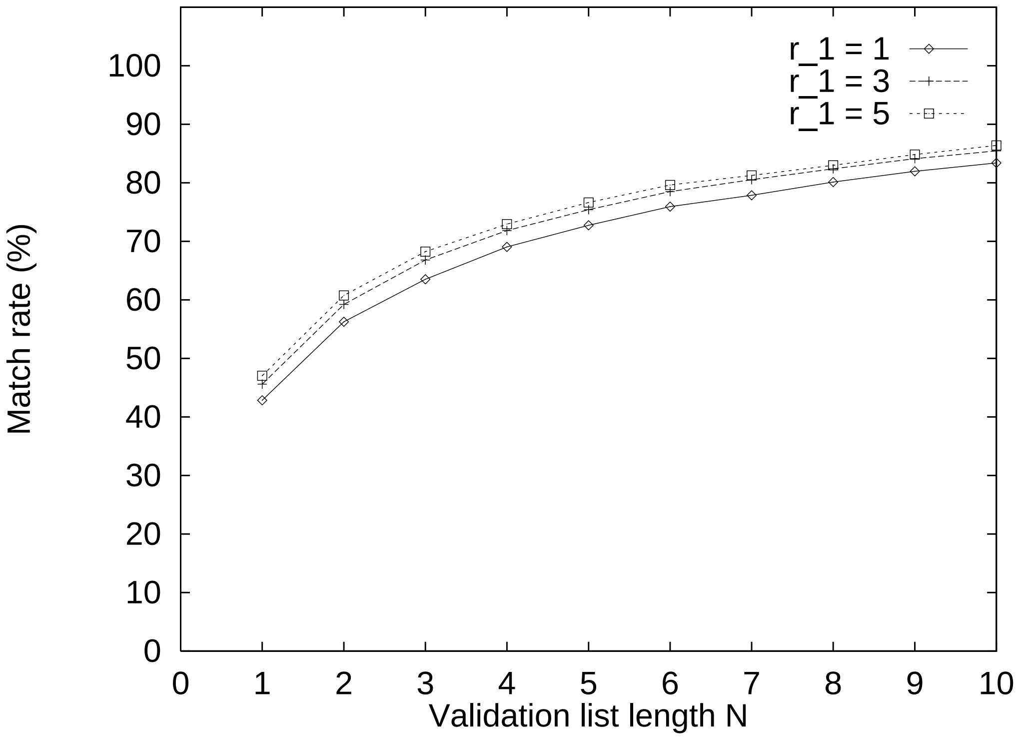

Figure 2: The match rate M^N is the probability that the the best keyword chosen by our method (G^1) for a map unit is found in the validation list of length N. The validation list of length N contains the N terms ranked best from independent lists of terms provided for the documents in the map unit. Each curve corresponds to a different radius r_1 of the "neutral zone", with the radius 5 producing the best results. For example, when the validation list length N was 2, and the radius r_1 was 5, the probability of a match was 61%. When the validation list consisted of all the terms provided with the documents in the map unit, the match rate was $95\% - 96\%$ for each of the radiuses.

map unit j be V_j^N, then a match m_j^N for map unit j was defined by

$$m_j^N = \begin{cases} 1 & \text{if } b_j \in V_j^N, \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The average match rate M^N was then the sum of m_j^N divided by the number of map units.

3.3 Validation results

The obtained match rates have been plotted in Fig. 2. The plot shows a considerably high match rate for the proposed method, in spite of the fact that the term lists used for comparison were probably intended for indexing for document retrieval purposes, whereas our method is intended for providing descriptors for document clusters.

Fig. 3 shows the topmost document map display labeled with the method described in section 2. Visual inspection and further exploration of the map confirm that the words selected as labels seem to be suitable descriptors of the respective areas.

4 Conclusions and discussion

We have proposed a method for selecting descriptor words for ordered clusterings or maps of textual data. The method can be used to provide labels or landmarks of graphical exploration interfaces. The variant G^0 is a general method for selecting keywords to characterize clusters of text. The variants G^1 and G^2 improve on G^0 by utilizing the ordering of the map. A fast approximation of the proposed method has been successfully applied for labeling a WEBSOM document map of seven million patent abstracts (Kohonen et al., 1999).

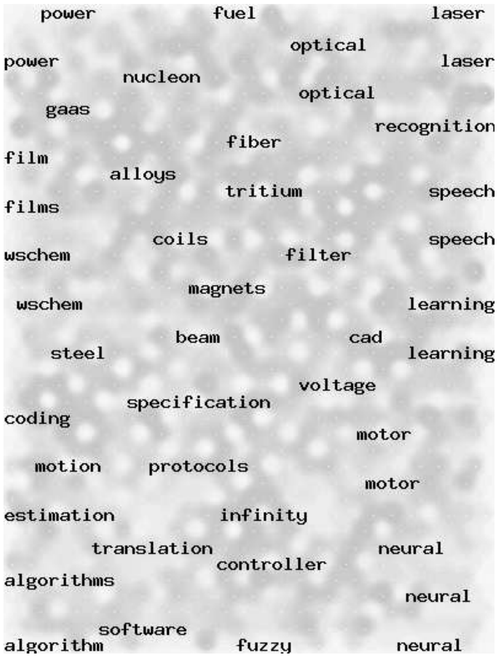

Figure 3: The top level map view of a document map that organizes a collection of $10,000$ INSPEC abstracts, labeled with the proposed labeling method described in Sec. 2.

So far the proposed method has proved to be extremely useful with exploration of text document maps. However, we believe that the method could provide valuable aid for interpretation and exploration of large collections of data in a wide variety of situations where the data of interest is numeric, but where some texts can be meaningfully associated with the data items.

5 Acknowledgements

We would like to thank INSPEC for the permission to use their database of scientific abstracts in our experiments.

References

Church, K. W. and Gale, W. A. (1995). Inverse document frequency (IDF): A measure of deviations from Poisson. In Yarowsky, D. and Church, K., editors, *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130. Massachusetts Institute of Technology, Cambridge, MA.

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996). Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). WEBSOM—self-organizing maps of document collections. *Neurocomputing*, 21:101–117.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Berlin, Heidelberg.

Kohonen, T., Kaski, S., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. (1999). Self organization of a massive text document collection. In Oja, E. and Kaski, S., editors, *Kohonen Maps*, pages 171–182, Amsterdam. Elsevier. In press.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1999). WEBSOM for textual data mining. *Artificial Intelligence Review*. In press.

Salton, G. and Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Technical Report 87-881, Cornell University, Department of Computer Science, Ithaca, NY.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.