16 Method for Characterizing Document Map Areas with Keywords

Krista Lagus and Samuel Kaski

When large collections of data are organized onto a map to visualize the collection, there is the need to characterize the different map areas. Characterization methods exist that can be used with any self-organizing maps (see Section 12), but with text document maps (see Section 14) there is a further possibility: keywords can be extracted from the documents and written on the map display to characterize the underlying area. The keywords aid in forming an overview of the document collection and ease interpretation of individual map areas. Furthermore, the keywords serve as navigation aids or *landmarks* during exploration of the map: they provide cues for maintaining a sense of location while moving along and across different zoom levels of the map display.

16.1 Keywords for map areas

A good descriptor of a cluster characterizes some outstanding property of the cluster in relation to the rest of the data collection. Therefore, when characterizing a cluster with a keyword, (1) the word must be outstanding within the cluster compared to other words in the cluster, and (2) the word must be relatively more outstanding in the cluster than elsewhere in the collection. These requirements can be combined into the following general form of a goodness measure G for word w in cluster k: $G(w,k) = F^{clust}(w,k) \times F^{coll}(w,k)$. By defining F^{clust} to describe the relative occurrence of word w in the cluster, and defining F^{coll} to relate the word to its occurrence in the rest of the collection, the obtained goodness measure implements our intuition of a good descriptor.

The same principle can be used to describe any map areas instead of clusters. However, the contents of map areas often change gradually without clear borders on the map. Therefore, instead of defining strict borders artificially, a better idea is to leave around each area to be characterized a "neutral zone" that neither supports nor inhibits a keyword (see Figure 26a). If the neutral zone is wide enough, it is probable that the topic clearly changes when moving over the zone. The goodness measure is described in detail in [1].

16.2 Labels for visual map displays

The keywords cannot be placed arbitrarily, however, since they may not overlap on the visual display, and since it is desirable that most areas obtain some characteristic label. Furthermore, the number of keywords should not be too large, to avoid overloading the visual display. To find for a map the optimal combination of N keywords, one would have to consider all the possible combinations of N keywords, which is in general prohibitively slow for larger maps. However, the following heuristic strategy may be used to obtain a good labeling in linear time: (1) Consider the proposed keywords in the *best-first order* determined by the goodness value G, and (2) accept a keyword as a label for the map display if the location associated with

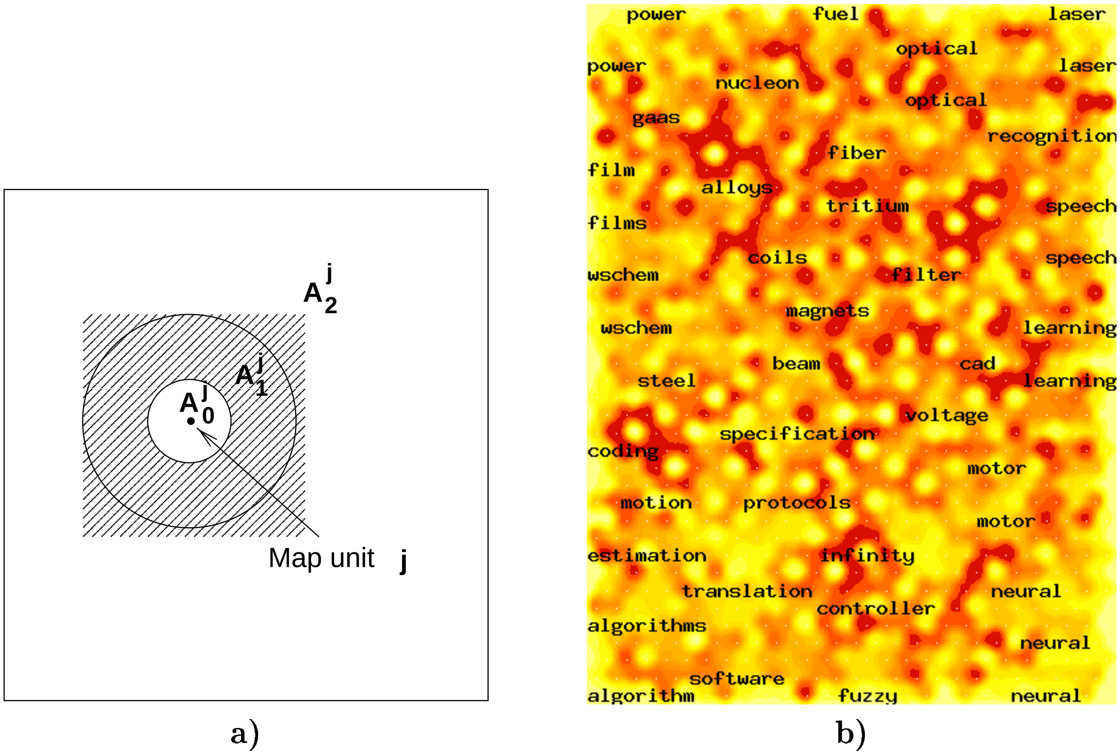

Figure 26: **a)** Only the white areas participate in calculating the goodness G of words in map unit j. The frequent occurrence of a word within the inner circle, area A_0^j, increases its possibility of becoming a keyword. Its frequent occurrence outside the outer circle, area A_2^j, inhibits the keyword. The shaded area A_1^j is disregarded and thus acts as a neutral zone neither giving support to the word nor inhibiting it. **b)** The top level map view of a document map that organizes a collection of $10,000$ scientific abstracts from INSPEC database, labeled with the described method.

the keyword is far enough from already accepted labels. Labelings with varying density can be obtained for different map display levels (zoom levels) by controlling the size of the map area in calculation of G as well as the minimum distance between accepted labels.

Using this scheme we have obtained satisfying labelings for WEBSOM document maps, e.g., the one in Figure 26b. Further examples of such maps can be explored starting from the WWW page `http://websom.hut.fi/websom/`.

The applicability of the method is not limited to document maps. We believe the method could be used to obtain characterizations for maps of very different kinds of data, given that there exists some text material that can be associated with the data items and therefore with the map units.

References

[1] K. Lagus and S. Kaski. *Keyword selection method for characterizing text document maps.* Submitted to ICANN'99, Ninth Int. Conf. on Artificial Neural Networks.