# A Visualization System for Social Information and Adaptive Clustering

Masamichi Iwamoto

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2001

**Keywords:** Document clustering, Visualization, Adaptive clustering, User feedback

The rapid diffusion of the Internet in recent years greatly changes our information environment. The quantity of information obtainable on the Internet becomes huge especially by World Wide Webs (WWW). In this information flooding environment, assisting systems of searching information are very important, which are now realized as search engines.

The search engines collect and rank the information or documents based on the user s request, usually expressed by keywords. The underlying assumption of these search engines is that there exists one document which facilitates the user s request. This is true for simple quiz type problems such as the population of Japan and the height of Mt. Fuji. However, difficult problems concerning social issues (environment, nuclear power plant, ...) cannot be solved in one document. Rather, various kinds of information should be collected from the user s viewpoint and used for deciding on one s position toward the problems.

The purpose of this system is to construct a system which assists the users to cope with such difficult problems. The functions of the systems are as follows: 1) clustering of the documents and displaying them using the terms, original texts, and the spatial relations of the terms on time axis, which helps the user with understanding how the problems have been discussed, and 2) re-clustering the document based on the user s feedback (we call this adaptive clustering ). The feedback consists of the user s selection of related documents and

the clustering mode: precision-based and recall-based. The former collects as few unrelated documents as possible; the latter collects as many related documents as possible. The system can learn the user s preference and display it as a cluster, which enables the user to realize her/his implicit thought.

The methods of adaptive clustering can be classified into two kinds: one is to change the term weight and the other is to increase the number of the terms. In this thesis, the term selection function, the rate of the term weight change, the number of the terms were used as parameters for evaluating the adaptive clustering methods. Term frequency-inverse document frequency (TFIDF), $\chi^2$ and the degree of context dependence were used as the term selection function. The rate of the term weight is changed from 1.5 to 10. The number of the term is changed from 10 to 100%. For the number of the terms, the inclusion of the terms which do not co-occur in the user s selected documents were also evaluated (we call this forced co-occurrence ).

The evaluation results showed that 1) TFIDF with 56.7% of the terms and the term weight change by 5 achieved the highest f-value, and 2) the forced co-occurrence with 41.7% of the terms and no term weight achieved the highest f-value, not degrading precision.