A New Efficient Clustering Algorithm for Organizing Dynamic Data Collection*

Kwangcheol Shin and Sangyong Han+

School of Computer Science and Engineering, Chung-Ang Univ.
221 HukSuk-Dong DongJak-Ku, Seoul, Korea
kcshin@archi.cse.cau.ac.kr, hansy@cau.ac.kr

Abstract. We deal with dynamic information organization for more efficient Internet browsing. As the appropriate algorithm for this purpose, we propose modified ART (artificial resonance theory) algorithm, which functions similarly with the dynamic Star-clustering algorithm but performs a more efficient time complexity of $O(nk)$, ($k << n$) instead of $O(n^2 log^2 n)$ found in the dynamic Star-clustering algorithm. In order to see how fast the proposed algorithm is in producing clusters for organizing information, the algorithm is tested on CLASSIC3 in comparison with the dynamic Star-clustering algorithm.

1 Introduction

From the very beginning of the information-oriented society era, gathering information has been very important issue. However, in current Internet environment there are too many documents; this makes users to waste time. Information organization techniques capable of automatically grouping related documents make it easy for users to recognize the contents of documents and to find what they want [1].

The most recent study on information organization deals with Star-clustering [2]. The Star-clustering algorithm presents the information system by applying the undirected, weighted similarity graph $G=(V,E,\omega)$ and forms a dense subgraph $G'=(V,E_\rho)$ based on G in order to organize the information. The Star-clustering algorithm also can be executed dynamically, which means each document clustered one by one. Compared to the formerly used average link or single link algorithm, the Star-clustering algorithm scored higher in the recall-precision measurement. However, in order to execute dynamic Star-clustering algorithm, the required time complexity is $O(n^2 log^2 n)$. And $O(n^2 log^2 n)$ is too mush time wasting when one has a massive amount of document groups to process in real time.

In this study, we suggest a new algorithm, which retains the benefits of the Star-clustering algorithm but has complexity only $O(kn)$, where k is the number of produced clusters. Our algorithm combines ART (artificial resonance theory) [3], a real

* Work supported by the ITRI of the Chung-Ang University.
+ Corresponding author.