Markov Cluster Shortest Path Founded upon the Alibi-breaking Algorithm

Jaeyoung Jung, Maki Miyake, and Hiroyuki Akama

Tokyo Institute of Technology, Department of Human System Science
2-12-1 O-okayama, Meguro-ku, Tokyo, 152-8552 Japan
{catherina, mmiyake, akama}@dp.hum.titech.ac.jp

Abstract. In this paper, we propose a new variant of the breadth-first shortest path search called Markov Cluster Shortest Path (MCSP). This is applied to the associative semantic network to show us the flow of association between two very different concepts, by providing the shortest path of them. MCSP is obtained from the virtual adjacency matrix of the hard clusters taken as vertices after MCL process. Since each hard cluster grouped by concepts as a result of MCL has no overlap with others, we propose a method called Alibi-breaking algorithm, which calculates the adjacency matrix of them in a way of collecting their past overlapping information by tracing back to the on-going MCL loops. The comparison is made between MCSP and the ordinary shortest paths to know the difference in quality.

1 Introduction

In the leading network science, the graph structure and scale problem has risen as a renewed matter of concern. The same thing is true of the corpus or cognitive linguistics that allows us to see the world of language as a large-scale graph of words. If a word is associated in a certain sense to the other, it is told that they are connected with each other and all the words taken in this way as nodes (vertices) are linked together by a set of edges corresponding here with the lexical association. In this structure, the shortest path between two random words or concepts represents their distance in semantic networks. Steyvers et al. (2003) showed that large-scale word association data possess a small-world structure characterized by the combination of highly clustered neighborhoods and a short average path length. According to them, the average shortest path (SP) length between any two words was 3.03 in the Undirected Associative Network of Nelson et al, 4.26 in their Directed Associative Network, 5.43 in Roget's thesaurus and 10.61 in WordNet.

It also held true in Ishizaki Associative Concepts Dictionary of Japanese Words (in abbreviation, ACD), which offered us lexical association data for graph manipulation. Its average shortest path (SP) length was 3.442 in the 43 word pairs randomly chosen from it. Despite such low values, however, it took a relatively long time (according to our experiment mentioned below, more than 1 minute on average) by the usual searching method that automatically traces the shortest routes based on the word node connectivity in semantic networks. This kind of word-to-word distance measure not