Deriving Tree-Structured Network Relations in Bibliographic Databases

Alisa Kongthon and Niran Angkawattanawit

Human Language Technology Laboratory
National Electronics and Computer Technology Center (NECTEC)
112 Thailand Science Park, Phahon Yothin Rd.
Klong Luang, Pathumthani, Thailand 12120
{alisa.kongthon,niran.angkawattanawit}@nectec.or.th

Abstract. This paper presents a new algorithm called "tree-structured networks" that can automatically construct parent-child (hierarchical structure) and sibling relationships (non-hierarchical structure) among concepts from a set of documents without use of data reduction or standard clustering techniques. The algorithm is applied to bibliographic databases such as INSPEC and EI Compendex toward the objective of enhancing research and development (R&D) management. Deriving tree-structured networks of research topics is an important goal in R&D management study. Parent-child relationships can help identify emerging areas in an existing field of research. Sibling relationships are interesting as well since they could represent interdisciplinary structures among related topical areas. Based on the initial testing on a set of publication abstracts, the proposed algorithm promises to offer richer structural information on relationships in text sources over the standard clustering techniques.

Keywords: Tree-structured networks, text mining, association rule mining, bibliographic databases, research and development management.

1 Introduction

Clustering text data in high-dimensional space is one of the most interesting topics among many other text mining applications. This paper presents the use of the object-oriented association rule mining (OOARM) technique [1] to automatically cluster related concepts and discern tree-structured networks in bibliographic databases such as INSPEC and EI Compendex. Tree-structured networks capture important aspects of both parent-child hierarchies (trees) and sibling relations (networks). It appears that most standard information retrieval and bibliometric analysis approaches using vector spaces or data reduction (e.g., Principal Components Analysis (PCA) or Latent Semantic Indexing) are able to identify relationships but not hierarchy.

2 The Proposed Algorithm

The proposed tree-structured network algorithm is implemented based on association rule mining. Kongthon et al (2007) introduced the new algorithm called Object-Oriented

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 504–505, 2007.

Association Rule Mining (OOARM) to effectively discover association rules from text data [1].

The basic tree-structured networks algorithm works as follows:

1. Find all *frequent term-clusters*
2. For each cluster, generate association rules with any other clusters
3. To obtain Parent-Child relations, find association rules that satisfy:
   $confidence(X \Rightarrow Y) = P(Y|X) \geq minParent$, where $minParent \leq 1$
   and
   $support(X) < \varepsilon * support(Y)$, where $0 < \varepsilon < 1$
   Y is then said to be 'parent' of X
4. To obtain sibling relations, find association rules between term clusters with ($minSibling \leq confidences < maxSibling$)
   where $0 < minSibling < maxSibling = minParent$

where

- X and Y are set of terms and $X \cap Y = \varnothing$
- An *itemset* is collection of one or more items
- Each *frequent term-cluster* is each *k*-frequent itemset that is generated by the OOARM algorithm
- Support is frequency of occurrence of an itemset
- Confidence of the rule X ⇒ Y is the conditional probability of Y given X.

3 Conclusion

In this paper, a tree-structured networks algorithm was proposed. The algorithm applies the association rule mining technique to discern conceptual relationships (parent-child and sibling) from text data sets. The results from the proposed algorithm were compared with the Principal Component Analysis (PCA) and the Hierarchical Agglomerative Clustering (HAC) approaches. Tree-structured networks promise to offer richer structural information on relationships in text information. Some remarkable features of tree-structured networks that are worth noticing include:

- Tree-structured networks do not require transformation of input data, instead it uses raw occurrences between input data.
- Parent-child and sibling relations can be derived from a set of documents without the use of data reduction or standard clustering techniques.

Reference

1. Kongthon, A., Mueller, R., Porter, A.L.: Object-Oriented Data Structured for Text Association Rule Mining. In: Proceedings of the 2007 Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) International Conference, pp. 1276–1279 (2007)