Finding Top-N Formal Concepts Guided by Dynamic Ordering of Objects

Li Ai XIANG, M. HARAGUCHI and Y. OKUBO
Graduate School of Infor. Sci. & Tech., Hokkaido University
E-mail: aixiang@kb.ist.hokudai.ac.jp, {mh, yoshiaki}@ist.hokudai.ac.jp

1 Introduction

One of the most fundamental issues in data mining has been the problem of enumerating formal concepts or closed patterns, and many advanced algorithms have been developed. In spite of their success, it is not an easy task to enumerate only a limited number of less frequent and more implicit concepts. A family of top-k algorithms has been proposed so as to generate only a limited number of frequent patterns, while the authors have also presented top-N algorithms to restrict the number of solutions, placing more emphasis on concept searches for less frequent and therefore closer to individual concepts. It is also intuitively clear that no one will like to have concepts that are too much individual. So, we try to maximize evaluation values of concepts under some constraints to exclude general and frequent ones.

2 Top-N algorithm

The top-N method solves the above problem by finding optimal solutions under some constraints. More precisely, we suppose two evaluation functions, $eval_{\mathcal{O}}$ and $eval_{\mathcal{F}}$, where $eval_{\mathcal{O}}(X)$ and $eval_{\mathcal{F}}(A)$ are the evaluations of extent X and intent A of a concept $< X, A >$, respectively. They are required to be increasingly monotonic w.r.t. set inclusion.

Objective: Enumerate every concept $< \psi A, A >$ with top N evaluation values $eval_{\mathcal{O}}(\psi A)$, where ψ, φ are functions defining the Galois connection. Moreover, $< \psi A, A >$ must be subject to

Length Constraint (required) $eval_{\mathcal{F}}(A) \geq \delta$ for excluding too frequent A, given a parameter $\delta > 0$.

Space Constraint (on demand):
POS: $S^+ \subseteq \psi A$ for an example object set S^+.
NEG: $S^- \cap \psi A = \phi$ for a negative object set S^-.
SUB: $K \subseteq A$ for a relevant feature set K.

Clearly, SUB provides an upper bound $< \psi K, \varphi\psi K >$ (the greatest concept in a sublattice). As the less frequent areas of concepts generally involve many number of maximal concepts under the length constraint, SUB provides a strong bias to find solutions even in those areas very quickly.

S^+ in POS and S^- in NEG work as positive and negative example sets. Particularly, S^+ defines a starting candidate closure, φS^+, in a bottom-up search algorithm. In the process of generating tentative closure A in a depth-first manner, an object x is called a candidate at A if $eval_{\mathcal{F}}(\varphi x \cap A) \geq \delta$. To avoid duplicated enumeration and to accelerate generating better extent w.r.t. its evaluation $eval_{\mathcal{O}}(\psi\varphi(\psi A \cup \{x\}))$, we arrange those candidates at each A as follows:

Dynamic Ordering $\preceq_A$: For candiates x, y at A, $x \prec_A y$ if $|\varphi x \cap A| < |\varphi y \cap A|$. As the feature set is smaller, more chances to imply another object.

Left Candidate : The tentative A has a history (path in the depth-first search tree) $x_1, ..., x_n$ from the initially given starting closure $A_0 = \varphi S^+$. Any candidate z at A_{j-1} s.t. $z \prec_{A_{j-1}} x_j$ is called a left candidate at A whenever $eval_{\mathcal{F}}(\varphi z \cap A) \geq \delta$.

Inverse Implication : When $A_n, x \rightarrow z$ for some left z at A_n, some preceding path stemming from $S^+, x_1, ..., x_{j-1}, z$ s.t. $z \prec_{A_{j-1}} x_j$ can generate the same closure. So we can safely cut off the branch x at A_n.

3 Experiments

The above duplication check and the dynamic ordering have been implemented in a standard branch-and-bound algorithm. For an incident relation with over 10,000 Web documents and about 1200 terms, we verified the effectiveness of the method. In a word, it succeeds within 10 seconds in detecting "crossover concept" connecting four categories of documents without depending on prior clustering or class information, given a few key words and some positive and negative documents.