Proc. 20th Int. Conf. on Data Enginnering, p. 835, Boston, MA, 2004, (Poster)

Efficient Similarity Search in Large Databases of Tree Structured Objects

Karin Kailing, Hans-Peter Kriegel, Stefan Schönauer
Institute for Computer Science
University of Munich
{kailing,kriegel,schoenauer}@informatik.uni-muenchen.de

Thomas Seidl
Institute for Computer Science
RWTH Aachen University
seidl@informatik.rwth-aachen.de

1. Introduction

Structured and semi-structured data are getting more and more important for modern database applications. Examples of such data include chemical compounds, CAD drawings, XML documents, web sites or image data. In addition to a variety of content-based attributes, complex objects mostly carry some kind of internal structure which often forms a hierarchy. Whereas for content information the concept of feature vectors has proven to be very successful, for the internal structure of the objects several similarity measures for trees have been proposed [4, 5]. However, the computational complexity of those measures limits their applicability to large databases. New efficient and effective lower-bounding filters for tree structured data in combination with a filter-refinement architecture [1, 3] can be used to overcome this problem.

2. Structural and Content-Based Filters

Filtering based on height of nodes. A successful way to filter unordered trees based on their structure is to take the height of nodes into account. A very simple technique is to use the height of a tree as a single feature. The difference of the height of two trees is an obvious lower bound for the edit distance between those trees, but this filter clearly is very coarse, as two trees with completely different structure but the same height cannot be distinguished. A more fine-grained and more sensitive filter is obtained by creating a histogram of node heights in a tree and using the difference between those histograms as a filter distance.

Filtering based on degree of nodes. The degrees of the nodes are another structural property of trees which can be used as a filter for the edit distances. Again, a simple filter can be obtained by using the maximal degree of all nodes in a tree as a single feature. The difference between the maximal degrees of two trees is an obvious lower bound for the edit distance. As before, this single-valued filter is very coarse, while using a degree histogram yields a more fine-grained filter criterion.

Content-Based Filters. Apart from the structure of the trees, the content features, expressed through node labels, have an impact on the similarity of attributed trees. The difference between the distribution of the values within a tree and the distribution of the values in another tree can be used to develop a lower-bounding filter. To ensure efficient evaluation of the filter, the distribution of those values has to be approximated for the filter step.

Combining different Filters. All of the above filters use a single feature of an attributed tree to approximate the edit distance. As the filters are not equally selective in each situation, several of the presented filters can be combined to further increase the filter selectivity.

3. Conclusion

We implemented our new approach for efficient similarity search in large databases of tree structures. Our experiments show that filtering significantly accelerates the complex task of similarity search for tree-structured objects. Moreover, they show that no single feature of a tree is sufficient for effective filtering, but only the combination of structural and content-based filters yields good results. More details can be found in [2].

References

[1] R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient similarity search in sequence databases.

[2] K. Kailing, H.-P. Kriegel, S. Schönauer, and T. Seidl. Efficient similarity search for hierachical data in large databases. In *Proc. EDBT 2004*, 2004.

[3] T. Seidl and H.-P. Kriegel. Optimal multi-step k-nearest neighbor search. In *Proc. ACM SIGMOD Int. Conf. on Managment of Data*, pages 154–165, 1998.

[4] S. Selkow. The tree-to-tree editing problem. *Information Processing Letters*, 6(6):576–584, 1977.

[5] K. Zhang. A constrained editing distance between unordered labeled trees. *Algorithmica*, 15(6):205–222, 1996.