Extended Abstract for the Intelligent Information Integration IJCAI'99 Workshop

Rim Al Hulou and **Amedeo Napoli** and **Emmanuel Nauer**
LORIA – UMR 7503
B.P. 239, 54506 Vandœuvre-les-Nancy Cedex, France
Email: {alhulou,napoli,nauer}@loria.fr

In this extended abstract, we briefly describe our current research works on semi-structured objects, polythetic classes and classification. These research works are carried out in the context of knowledge discovery in databases (KDD), text mining and scientific and technical information analysis, especially in biological, chemical, and medical databases. In these application domains, many problems arise from the heterogeneity and the irregularity of data. For example, the exchange of files coding molecules may be very difficult and awkward since there are more than fifty different file formats available. Thus, the communications between different databases –and different knowledge bases– have to be facilitated and to be made more practical. This is one of the reasons that lead us to the study of formalisms taking into account semi-structured data [Abiteboul,1997] [Buneman *et al.*,1997]. Semi-structured data are data whose structure is not regular, heterogeneous, partial, has not a fixed format and quickly evolves. These characteristics are typical of data available on the Web, of data coming from heterogeneous information sources, of multimedia data, etc. However, research works on semi-structured data may be also exploited for combining knowledge-based systems with databases management systems, especially in the context of very large knowledge bases and databases, a crucial need for information systems at present.

In order to handle semi-structured data, knowledge discovery techniques –and particularly classification techniques– can be used to discover regularities in large amount of data. These regularities can then be used as knowledge units in a knowledge-based system [Fayyad *et al.*,1996]. Text mining and information analysis can be seen as two kinds of KDD processes, leading both to take into account text files with different formats, e.g. files on the Web. Text mining and information analysis are expected to provide text syntheses that can considered as knowledge sources from which knowledge units can be extracted. Reasoning procedures can then be applied to these knowledge units for text understanding, query assistance, technological survey, and also to design, maintain and extend a corporate memory (integration of data and knowledge coming from different sources to build the corporate memory).

In order to represent and classify semi-structured data, we rely on the the formalism of semi-structured objects introduced in [Bertino *et al.*,1999]. Semi-structured objects are objects –in the sense of object-oriented programming or object-based representation systems– that can be created without any class declaration and described with disjunction of attributes, and that can be then classified, according to their description, in a hierarchy of reference classes ; this hierarchy plays the role of th schema in object-oriented databases. Moreover, we draw a parallel between semi-structured objects and polythetic classes, i.e. a classification formalism that has been introduced many years ago in the field of data analysis to take into account the irregularity of real-world concepts, mainly biological concepts, a natural kind of semi-structured data.

More precisely, a class $\mathtt{C}$ defined by a set of attributes $\mathtt{A} = \{\mathtt{a}_1, ..., \mathtt{a}_n\}$ is said to be *polythetic* if and only if (i) every object that is an instance of the class $\mathtt{C}$ has an "important" number of attributes of $\mathtt{A}$, (ii) every attribute of $\mathtt{A}$ belongs to an "important" number of instances of $\mathtt{C}$, (iii) there is not necessarily an attribute of $\mathtt{A}$ belonging to every instance of $\mathtt{C}$. Thus, a polythetic class $\mathtt{C}$ can be formalized as $\mathtt{C} = \mathtt{a}_1 \sqcap \mathtt{a}_2 \sqcap ... \sqcap \mathtt{a}_n$, where $\mathtt{a}_i$ denotes an attribute ; some of the $\mathtt{a}_i$ are "simple" or primitive attributes, i.e. (not decomposable attributes, while some of the $\mathtt{a}_i$ can be of the form $\mathtt{a}_i = \mathtt{b}_{i_1} \sqcup \mathtt{b}_{i_2} \sqcup ... \sqcup \mathtt{b}_{i_n}$, where $\mathtt{b}_{i_j}$ denotes primitive attributes.

The classification process of a semi-structured object $\mathtt{x}$ depends on a search for the most specific subsumer of $\mathtt{x}$ in a reference hierarchy of (polythetic) classes (this hierarchy defines a model of the domain of the data). The classification process relies on the following constraints: The number of attributes of the object $\mathtt{x}$ to be classified must be (at least) greater than or equal to the classification threshold $\mathtt{T(C)}$ of the candidate class $\mathtt{C}$ ($\mathtt{T(C)}$ fixes the minimal number of attributes that $\mathtt{x}$ must have to be a member of the extension of $\mathtt{C}$). The values of the attributes of the object $\mathtt{x}$ must be conform to the corresponding attribute specifications declared in the candidate class $\mathtt{C}$. The class $\mathtt{C}$ with the minimal number of additional attributes with respect to $\mathtt{x}$ is the best candidate reference class.

Reasoning on semi-structured object is mainly based query answering and reasoning on queries, e.g. recognizing that a query $\mathtt{Q}_1$ is more general than a query $\mathtt{Q}_2$ or that a query $\mathtt{Q}_1$ is analogous to a query $\mathtt{Q}_2$. A query $\mathtt{Q}_1$ is more general than a query $\mathtt{Q}_2$ if the class representing the query $\mathtt{Q}_1$ subsumes the class representing the query $\mathtt{Q}_2$ [Buchheit *et al.*,1994]. A query $\mathtt{Q}_1$ is analogous to a query $\mathtt{Q}_2$ if there exists a *similarity path* between the class representing $\mathtt{Q}_1$ and the class representing $\mathtt{Q}_2$ in $\mathcal{H}$. Similarity paths have been introduced for adaptation in case-based reasoning [Lieber and Napoli,1998] and are reused in the context of query comparison.

References

[Abiteboul, 1997] S. Abiteboul. Querying Semi-Structured Data. In F. Afrati and P. Kolaitis, editors, *Database Theory – ICDT'97, 6th International Conference, Delphi, Greece*, Lecture Notes in Artificial Intelligence 1186, pages 1–18. Springer, Berlin, 1997.

[Bertino *et al.*, 1999] E. Bertino, G. Guerrini, I. Merlo, and M. Mesiti. An approach to classify semi-structured objects. In *To appear in Proceedings of ECOOP'99*. Springer, Berlin, 1999.

[Buchheit *et al.*, 1994] M. Buchheit, M.A. Jeusfeld, W. Nutt, and M. Staudt. Subsumption between queries to object-oriented databases. *Information Systems*, 19(1):33–54, 1994.

[Buneman *et al.*, 1997] P. Buneman, S. Davidson, M. Fernandez, and S. Suciu. Adding Structure to Unstructured Data. In F. Afrati and P. Kolaitis, editors, *Database Theory – ICDT'97, 6th International Conference, Delphi, Greece*, Lecture Notes in Artificial Intelligence 1186, pages 336–330. Springer, Berlin, 1997.

[Fayyad *et al.*, 1996] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From Data Mining to Knowledge Discovery: An Overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press / MIT Press, Menlo Park, California, 1996.

[Lieber and Napoli, 1998] J. Lieber and A. Napoli. Correct and Complete Retrieval for Case-Based Problem-Solving. In H. Prade, editor, *Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98), Brighton, UK*, pages 68–72. John Wiley & Sons Ltd, Chichester, 1998.