Objets semi-structurés, classes polythétiques et classification

Rim Al Hulou, Amedeo Napoli, Emmanuel Nauer

Objets semi-structurés, classes polythétiques et classification

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

Résumé

Dans cet article, nous présentons un travail de recherche en cours de développement sur la représentation et la manipulation de données semi-structurées, dans le contexte des systèmes de représentation de connaissances par objets (RCO). Nous nous appuyons sur la notion d'objet semi-structuré qui peut être défini comme un objet sans classe, intégrant des disjonctions d'attributs. Un tel objet peut être classifié dans une hiérarchie de classes de référence qui représente la connaissance sur un domaine donné, en vue de mener à bien des raisonnements et résoudre des problèmes. La prise en compte d'objets semi-structurés conduit à considérer les classes de référence comme des classes polythétiques, au sens où elles sont définies par une combinaison de disjonctions et de conjonction d'attributs. Ce travail de recherche essaie également de faire le lien entre le traitement des données semi-structurées dans le cadre des systèmes de RCO et le traitement des classes polythétiques en analyse de données ou en apprentissage. Les cadres d'applications de ce travail de recherche sont multiples : extraction de connaissances dans les bases de données, fouille de textes, intégration et croisement de données hétérogènes, conception et couplage de grandes bases de connaissances et de grandes bases de données.

Mots-clés données et objets semi-structurés, classe polythétique, représentation de connaissances par objets, classification.

1 Introduction

Dans cet article, nous décrivons un travail de recherche en cours qui met en relation le traitement des données semi-structurées, les objets semi-structurés, les classes polythétiques et la classification. Ce travail de recherche entre dans le cadre de travaux sur l'extraction de connaissances dans des bases de données (ECBD), la fouille de textes pour l'analyse de l'information scientifique et technique, l'intégration et le croisement de données hétérogènes. Des problèmes complexes d'intégration et de croisement de données hétérogènes se rencontrent particulièrement en chimie, biologie ou médecine : par exemple, il existe plus d'une cinquantaine de formats différents pour coder des molécules, ce qui rend les échanges de code de molécules très difficiles, nuisant ainsi à la portabilité des programmes de manipulation de molécules [Campagne,1998]. Les communications entre différentes bases de données, ainsi qu'entre bases de données et bases de connaissances doivent être rendues plus pratiques et plus naturelles. C'est là une des raisons de l'intérêt porté actuellement sur les formalismes permettant de prendre en compte des données semi-structurées [Abiteboul,1997]. Des données semi-structurées sont des données hétérogènes, non régulières, sans format fixe bien déterminé, et évoluant rapidement. De telles caractéristiques sont typiques par exemple de données en provenance du web ou de données multimédia, etc. Les

travaux sur les données semi-structurées revêtent aussi une certaine importance dans la conception de très grandes bases de données et de très grandes bases de connaissances. De telles bases sont des ingrédients essentiels des systèmes d'informations futurs.

Pour prendre en compte des données semi-structurées, des techniques d'ECBD [Fayyad *et al.*,1996] — et en particulier des techniques d'analyse de données et de classification conceptuelle — peuvent être utilisées pour découvrir des régularités dans les données et faire émerger une *structure primitive*. Ces régularités peuvent, le cas échéant, se transformer en éléments de connaissances et être utilisées dans le cadre d'un système à base de connaissances pour résoudre un problème sur un domaine d'étude donné. Par exemple, la fouille de textes et l'analyse de l'information scientifique et technique peuvent être vues comme des processus d'ECBD, qui conduisent tous deux à prendre en compte des fichiers textes hétérogènes, comme les fichiers en provenance du web. La fouille de textes et l'analyse de l'information doivent permettre de fournir des synthèses de textes, qui peuvent ensuite être considérées comme des sources potentielles de connaissances à partir desquelles peuvent être extraites certaines unités de connaissances. Des procédures de raisonnement peuvent ensuite être appliquées à ces unités de connaissances pour répondre à des questions concernant les textes, en vue de les comprendre, les manipuler, les résumer, les indexer (pour des besoins de veille technologique), pour concevoir, maintenir et faire évoluer des mémoires d'entreprise (l'intégration de bases de donnés hétérogènes est un des premiers problèmes dans ce cadre).

Dans la suite, nous décrivons une approche qui permet de représenter et de classifier des données semi-structurées dans le cadre des systèmes de RCO. Les éléments de base de cette approche sont les objets semi-structurés, qui sont des objets — au sens de la programmation par objets ou des systèmes de RCO — qui peuvent être créés sans avoir de classe d'intanciation, comme des prototypes [Dony *et al.*,1998] ou des *frames* [Euzenat,1998], et qui, de plus, peuvent être décrits par des disjonctions d'attributs. Ces objets « orphelins » peuvent ensuite être reclassés par rapport à une hiérarchie de référence, qui joue un rôle analogue à celui tenu par le schéma des classes dans les bases de données à objets. De plus, nous mettons en évidence le parallèle qui existe entre les objets semi-structurés et les classes polythétiques : le formalisme des classes polythétiques a été introduit pour prendre en compte les irrégularités des données du monde réel, comme les données de la biologie [Lebbe,1991] [Vignes,1991].

Le papier est organisé de la façon suivante. En premier lieu, nous introduisons l'environnement des systèmes de RCO dans lequel ce travail de recherche est développé. Ensuite, nous présentons les objets semi-structurés et mettons en évidence les liens qui existent entre ces objets d'un type particulier et les classes polythétiques. Enfin, nous discutons de classification et de problèmes de raisonnements avec des objets semi-structurés et nous montrons les perspectives de travail associées.

2 Objets semi-structurés et classes polythétiques

2.1 Les systèmes de RCO

Nous nous plaçons dans le cadre des systèmes de *représentation de connaissances par objets* — ou systèmes de RCO — [Simon *et al.*,1998] : un tel système s'appuie sur une *hiérarchie* de classes notée $\mathcal{H} = (\mathcal{X}, \omega, \sqsubseteq)$, où les classes dans $\mathcal{X}$ sont reliées entre elles par une relation de *subsomption* $\sqsubseteq$, ω étant la racine de la hiérarchie (l'élément maximum pour $\sqsubseteq$). Une classe représente un concept du monde réel et se compose d'un ensemble d'attributs qui forme l'*intension* de la classe et qui décrit les caractéristiques et le comportement du concept représenté. Une classe peut être instanciée pour produire un ensemble d'*instances* — les objets individuels —, qui forme

l'*extension* de la classe. Dans l'approche considérée, un objet peut également être créé sans classe de référence, tout comme un *prototype* dans un langage de prototypes [Dony *et al.*,1998] : dans ce cas, l'objet est considéré comme une instance de ω, la racine de $\mathcal{H}$.

Les systèmes de RCO sont utilisés pour représenter des connaissances sur un domaine donné et pour mener des raisonnements sur ce domaine, en vue d'y résoudre des problèmes. La hiérarchie $\mathcal{H}$ correspond à la base de connaissances du système. Le raisonnement s'appuie principalement sur l'*héritage* de propriétés dans la hiérarchie $\mathcal{H}$ et sur le processus de *classification*. L'héritage permet de gérer la circulation et le partage de propriétés dans la hiérarchie — les sous-classes héritent les caractéristiques de leurs super-classes — tandis que la classification permet d'insérer un nouvel élément, classe ou instance, dans la hiérarchie, mais aussi de gérer des *requêtes* effectuées sur la hiérarchie pour déduire de nouvelles informations [Borgida et McGuinness,1996]. De plus, le processus de classification peut être appréhendé comme un processus de « classification conceptuelle » au sens où il permet de classifier un objet sans classe dans l'extension de la classe la plus appropriée de la hiérarchie $\mathcal{H}$; l'expression « classification conceptuelle » fait ici référence à un processus dynamique et éventuellement incrémental de construction et d'organisation de classes décrivant des concepts du monde réel [Langley,1996].

2.2 Les classes polythétiques

Pour prendre en compte et manipuler des données semi-structurées et mener à bien des raisonnements sur de telles données, par exemple pour des besoins de résolution de problèmes ou d'ECBD, nous nous appuyons sur les notions de classes polythétiques et d'objets semi-structurés. Les classes de la hiérarchie $\mathcal{H}$ ne sont plus considérées comme des classes *monothétiques* — la possession pour un individu $\mathtt{x}$ d'un ensemble fixé d'attributs détermine une condition nécessaire et suffisante pour décider l'appartenance de $\mathtt{x}$ à la classe — mais comme des classes *polythétiques* [Sokal et Sneath,1963] [Lerman,1970] [Jardine et Sibson,1971].

Un classe $\mathtt{C}$ définie par un ensemble d'attributs $\mathtt{A} = \{\mathtt{a}_1, ..., \mathtt{a}_n\}$ est dite *polythétique* si et seulement si :

- Tout objet qui est une instance de la classe $\mathtt{C}$ possède un nombre « important » — mais non nécessairement fixé — d'attributs de $\mathtt{A}$.
- Tout attribut de $\mathtt{A}$ appartient à un nombre « important » d'instances de $\mathtt{C}$.
- Il n'existe pas nécessairement un attribut de $\mathtt{A}$ qui appartienne à chaque instance de $\mathtt{C}$.

Ainsi, une classe polythétique n'est plus uniquement une conjonction d'attributs mais peut intégrer des disjonctions d'attributs. Plus précisément, une classe $\mathtt{C}$ peut se décrire par une conjonction $\mathtt{C} = \mathtt{a}_1 \sqcap \mathtt{a}_2 \sqcap ... \sqcap \mathtt{a}_n$, où $\mathtt{a}_i$ dénote un attribut ; certains attributs $\mathtt{a}_i$ sont « simples » ou « primitifs » au sens où ils ne sont pas décomposables, tandis que certains $\mathtt{a}_i$ sont *décomposables* ou *disjonctifs*, c'est-à-dire de la forme $\mathtt{a}_i = \mathtt{b}_{i_1} \sqcup \mathtt{b}_{i_2} \sqcup ... \sqcup \mathtt{b}_{i_n}$, où les $\mathtt{b}_{i_j}$ dénotent des attributs primitifs.

Cette façon d'appréhender les classes est à rapprocher de la définition du types de données *Union* dans les bases de données à objets [Abiteboul *et al.*,1997] [Guerrini *et al.*,1998], des concepts définis avec le constructeur *Or* dans les logiques de descriptions [Napoli,1998], mais aussi des points de vue dans les systèmes de RCO [Euzenat,1998]. De cette façon, les irrégularités des données semi-structurées peuvent être appréhendées plus naturellement.

2.3 Objets semi-structurés

En nous appuyant sur les idées et le formalisme introduit dans [Bertino *et al.*,1999], nous définissons un *objet semi-structuré* comme un objet sans classe, instance de la racine de la hiérarchie

ω ; il faut noter qu'ici ω joue un rôle similaire au type `spring` des objets sans classes dans [Bertino *et al.*,1999]. La hiérarchie $\mathcal{H}$ de référence est une hiérarchie de classes polythétiques qui sert de *modèle* du domaine relatif aux données. Ce modèle joue en fait le rôle de *guide* de compréhension des données — *data guide* dans [Abiteboul,1997] — qui fournit une structure primitive sur laquelle s'appuyer pour manipuler des données semi-structurées. En particulier, représenter des données semi-structurées sous forme d'objets semi-structurés puis les classifier par rapport à la hiérarchie des classes polythétiques est une façon de découvrir des régularités dans de telles données. Ainsi, les données semi-structurées, par l'intermédiaire des objets semi-structurés, peuvent être représentées par des objets, regroupées en classes hiérarchisées, pour être utilisées dans des raisonnements.

Le processus de classification d'un objet semi-structuré `x` se déroule de la façon suivante (pour simplifier, nous supposons qu'un objet semi-structuré ne peut avoir qu'une seule classe de rattachement). Une recherche des classes subsumantes les plus spécifiques de `x` est effectuée en profondeur dans la hiérarchie $\mathcal{H}$ et doit obéir aux contraintes globales suivantes :

- Les valeurs des attributs de l'objet `x` doivent être conformes aux spécifications des valeurs d'attributs introduites dans la classe candidate `C`.
- Pour un attribut disjonctif $\mathtt{a_i} = \mathtt{b_{i_1}} \sqcup \mathtt{b_{i_2}} \sqcup ... \sqcup \mathtt{b_{i_n}}$, la classe candidate `C` possédant le nombre minimal d'attributs primitifs $\mathtt{b_{i_k}}$ par rapport à l'objet à classifier `x` est celle qui est choisie comme classe de référence. Il faut noter ici une différence avec la classification standard où c'est la classe la plus spécifique qui est recherchée, donc celle qui possède le plus d'attributs ; ici, au contraire, c'est la classe possédant le moins d'attributs dans une disjonction qui est choisie.

Si ces contraintes globales sont vérifiées par l'objet semi-structuré `x` par rapport à une classe candidate `C`, alors `C` devient la classe polythétique de référence pour l'objet `x`. Lorsque plusieurs classes de référence sont candidates selon les mêmes critères, une classe de référence est choisie arbitrairement.

Il est possible d'introduire des variations dans le processus de classification esquissé ci-dessus. Ainsi, pour marquer le caractère polythétique des classes de la hiérarchie, un *seuil de classification*, noté $\mathtt{T(C)}$, peut être associé avec une classe polythétique `C`, pour fixer le nombre minimal d'attributs qu'un objet particulier `x` doit posséder pour un attribut disjonctif. Dans le même ordre d'idées, un degré de *conformité* et un degré d'*hétérogénéité* sont introduits dans [Bertino *et al.*,1999]. Le degré de conformité mesure la similarité entre le type d'un objet semi-structuré et le type de la classe de référence. Intuitivement, le degré de conformité mesure l'importance du nombre d'attributs que l'objet semi-structuré possède par rapport à la classe polythétique de référence. Le degré d'hétérogénéité mesure l'hétérogénéité de l'extension d'une classe polythétique. Intuitivement, le degré d'hétérogénéité mesure la proportion d'attributs qui sont possédés par le plus grand nombre d'objets.

2.4 Raisonner avec des objets semi-structurés

Les données semi-structurées peuvent être représentées par des objets semi-structurés et ensuite interprétées par rapport à leur classe polythétique de référence pour différents besoins : raisonnement et résolution de problèmes, satisfaction de requêtes (une requête `Q` est représentée par une classe et satisfaire `Q` revient à classifier cette classe dans la hiérarchie), raisonnements sur les requêtes comme la classification de requêtes ou la recherche de requêtes analogues dans le cadre du raisonnement à partir de cas, etc. En particulier, une requête $\mathtt{Q_1}$ est plus générale qu'une requête $\mathtt{Q_2}$ si la classe représentant la requête $\mathtt{Q_1}$ subsume la classe représentant la requête $\mathtt{Q_2}$ [Buchheit

et al.,1994] [Levy et Suciu,1997]. Une requête $\mathtt{Q}_1$ est *analogue* à une requête $\mathtt{Q}_2$ s'il existe un *chemin de similarité* entre la classe représentant $\mathtt{Q}_1$ et celle représentant $\mathtt{Q}_2$ dans $\mathcal{H}$. Les chemins de similarités ont été introduits en raisonnement à partir de cas pour guider la remémoration d'un cas en fonction de l'adaptation de ce cas en vue de résoudre un nouveau problème [Lieber et Napoli,1998] [Simon *et al.*,1998]. Les chemins de similarité peuvent être réutilisés dans le cadre de la comparaison de requêtes. Pour simplifier, nous considérons que de tels chemins sont composés uniquement de spécialisations et de généralisations, autrement dit qu'un chemin de similarité peut exister entre deux classes si les deux classes ont un ancêtre commun dans la hiérarchie $\mathcal{H}$ autre que ω.

2.5 Conclusion

Dans cet article, nous avons brièvement présenté un travail de recherche en cours de développement, qui essaie de mettre en relation les études faites sur les objets semi-structurés et sur les classes polythétiques. Dans notre approche, nous supposons qu'il existe une hiérarchie de classes représentant des connaissances sur le domaine étudié — un modèle de ce domaine — où les classes sont considérées comme des classes polythétiques, de telles classes pouvant combiner des disjonctions et des conjonctions d'attributs. Des données semi-structurées en rapport avec le domaine étudié peuvent être représentées par des objets qui sont sans classe *a priori* et qui peuvent être ensuite classifiés par rapport aux classes du modèle du domaine. Il est possible alors d'exploiter de telles données à l'aide de procédures de raisonnement pour résoudre des problèmes sur le domaine étudié.

Cette approche peut en particulier être utilisée dans la manipulation intelligente de données, et en particulier dans le cadre de l'extraction de connaissances dans des bases de données, dans l'intégration de bases de données hétérogènes ou encore dans le couplage de bases de données et de bases de connaissances pour la conception de mémoire d'entreprises.

Références

[Abiteboul *et al.*, 1997] S. Abiteboul, S. Cluet, V. Christophides, T. Milo, G. Moekotte et J. Siméon. Querying Documents in Object Databases. *International Journal on Digital Libraries*, 1997.

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

[Bertino *et al.*, 1999] E. Bertino, G. Guerrini, I. Merlo et M. Mesiti. An approach to classify semi-structured objects. Dans *Proceedings of ECOOP'99, Lisboa (Portugal)*, rédacteur R. Guerraoui, Lecture Notes in Computer Sciences 1628, pages 416–440. Springer, Berlin, 1999.

[Borgida et McGuinness, 1996] A. Borgida et D.L. McGuinness. Asking Queries about Frames. Dans *Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning (KR'96), Cambridge, Massachusetts*, pages 340–349, 1996.

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

[Campagne, 1998] F. Campagne. *Conception et développement de systèmes d'aide à la compréhension de données biologiques et moléculaires*. Thèse de chimie informatique et théorique, Université Henri Poincaré, Nancy, 1998.

[Dony *et al.*, 1998] C. Dony, J. Malenfant et D. Bardou. Les langages à prototypes. Dans *Langages et modèles à objets — État des recherches et perspectives*, rédacteurs R. Ducournau, J. Euzenat, G. Masini et A. Napoli, Collection Didactique D-019, pages 227–256. INRIA, Le Chesnay, 1998.

[Euzenat, 1998] J. Euzenat. Représentation de connaissances par objets. Dans *Langages et modèles à objets — État des recherches et perspectives*, rédacteurs R. Ducournau, J. Euzenat, G. Masini et A. Napoli, Collection Didactique D-019, pages 293–319. INRIA, Le Chesnay, 1998.

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

[Guerrini *et al.*, 1998] G. Guerrini, E. Bertino et R. Bal. A formal definition of the chimera object-oriented data model. *Journal of Intelligent Information Systems*, 11:5–40, 1998.

[Jardine et Sibson, 1971] N. Jardine et R. Sibson. *Mathematical Taxonomy*. John Wiley & Sons, London, 1971.

[Langley, 1996] P. Langley. *Elements of Machine Learning*. Morgan Kaufmann Publishers, San Francisco, California, 1996.

[Lebbe, 1991] J. Lebbe. *Représentation des concepts en biologie et médecine (Introduction à l'analyse des connaissances et à l'identification assistée par ordinateur)*. Thèse d'Informatique, Université Pierre et Marie Curie (Paris 6), 1991.

[Lerman, 1970] I.C. Lerman. *Les bases de la classification automatique*. Gauthier-Villars Éditeurs, Paris, 1970.

[Levy et Suciu, 1997] A.Y. Levy et D. Suciu. Deciding Containment for Queries with Complex Objects. Dans *Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Tucson, Arizona*, 1997.

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

[Napoli, 1998] A. Napoli. Une introduction aux logiques de descriptions. Dans *Langages et modèles à objets — État des recherches et perspectives*, rédacteurs R. Ducournau, J. Euzenat, G. Masini et A. Napoli, Collection Didactique D-019, pages 321–350. INRIA, Le Chesnay, 1998.

[Simon *et al.*, 1998] A. Simon, A. Napoli, J. Lieber et A. Ketterlin. Aspects de la classification dans un système de représentation des connaissances par objets. Dans *Actes des sixièmes rencontres de la Société Francophone de Classification (SFC'98), Montpellier*, rédacteurs O. Gascuel et G. Caraux, pages 205–209. Publication de l'École d'Agronomie de Montpellier, 1998.

[Sokal et Sneath, 1963] R.R. Sokal et P.H.A. Sneath. *Principles of Numerical Taxonomy*. Freeman, San Francisco (CA), USA, 1963.

[Vignes, 1991] R. Vignes. *Caractérisation automatique de groupes biologiques*. Thèse d'Informatique, Université Pierre et Marie Curie (Paris 6), 1991.