Contextual Data and Domain Knowledge for Incorporation in Knowledge Discovery Systems

Alex G. Büchner[1], John G. Hughes[1] and David A. Bell[2]

[1] Northern Ireland Knowledge Engineering Laboratory, University of Ulster
[2] School of Information and Software Engineering, University of Ulster
Shore Road, Newtownabbey, BT37 0QB, UK
phone: +44 (0)1232 368394 fax: +44 (0)1232 366068
email: {ag.buchner, jg.hughes, da.bell}@ulst.ac.uk

Abstract. The concepts of contextual data and domain knowledge is proposed and incorporated in a generic knowledge discovery architecture. The outlined concepts are supported electronic commerce examples.

Introduction

Data and domain knowledge are the two most essential input ingredients for data mining applications. The former is either provided by operational databases or by their warehoused counterparts in form of materialised views, which can be reused for multiple model building exercises. The latter, in whatever form provided, has to be re-specified, or at least modified, depending in which context patterns are to be discovered. This limits the concept of domain knowledge, and thus, the objective of this paper is to propose the notion of contextual data and domain knowledge, which can be reused across multiple related knowledge discovery exercises.

The structure of this paper is as follows. First, the organisation of contexts is outlined, which includes a brief recapitulation of contextual data as well as contextual domain knowledge. Then, the concepts, which are based on previous work, are incorporated in a generic knowledge discovery framework.

Organisation of Contexts

A context represents behavioural aspects that are shared by attributes of the same ontology, which are organised hierarchically. The general idea is that every single attribute instance is being allotted an additional attribute context identifier in a multi-database scenario, where each attribute instance is represented by a semantic value [1]. The same concept is also applied to domain knowledge, which guarantees consistent handling of the two components.

Contextual Data

Contextual data is represented by semantic values, which consist of a type, a value and a context [2]. This approach is a generalisation of proposals, in which every attribute — but not each attribute instance — has a specific context allotted to it. In order to allow database operations function correctly, contextual comparison operations have been introduced, which are handled by a data context mediator. The mediator is an evaluation mechanism that takes two semantic values and returns their type- *and* context-specific order. A variety of constructs has been suggested in the literature, but to reconcile context heterogeneity for data mining input, functions (which encompass arithmetic expressions as well as rules) and tables are sufficient.

```
context mediator ß s1, s2
        func <conversion_function> | table <collation>
context mediator à {-1, 0, 1}
```

Fig. 1. Data Context Mediator Structure

The data context mediator, which has been embedded in an object data model, supports atomic, abstract, as well as complex data types (see [2] for more details).

Contextual Domain Knowledge

Domain knowledge can be used for making patterns more visible, for constraining the search space, for finding more accurate knowledge, and for filtering out uninteresting patterns [3]. For the purpose of a contextual data mining environment, four types of domain expertise are supported: Taxonomies, which encompass bandings (ranges), concept hierarchies and network models, constraints (aka. attribute-relationship rules), previously discovered knowledge, and user preferences (see [3] for details). Each domain knowledge type is investigated from two different dimensions. The first is concerned with the degree of reality, where reality is represented in a spectrum from a physical world to a logical model world (Fig. 2.). The second is interested in the degree of reusability of the specified types of domain knowledge [3].

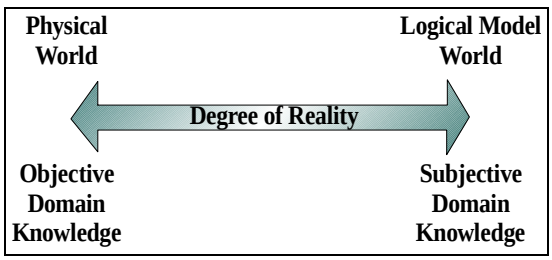

Fig. 2. Domain Knowledge Degrees of Reality

Objective domain knowledge consists of a set of quasi facts within the domain a data mining exercise is carried out. Although it can have a certain degree of context-dependency, it is almost always kept as holos and only exchanged in total for its

contextual counterpart. *Subjective domain knowledge* has a higher degree of context-dependency than its objective counterpart. As a consequence, either entire domain knowledge entities or large parts thereof have to exist for multiple contexts. It is desirable to handle the entire range of domain knowledge degrees of reality using the same underlying techniques. Thus, contextual domain knowledge has been proposed, independent of its degree of reality. Basically, each piece of domain knowledge is allotted a context, which conforms to the handling of contextual data.

For example, contextual concept hierarchies are of high interest to data mining exercises. Because each node in a hierarchy is allotted a context identifier, it is possible to reuse a collection of nodes and their sub-trees. The marketing manager Europe (context c_m) might only be interested in the sub-hierarchy with all European countries, whereas the person responsible for introducing a product at educational level (c_e) would only be concerned about the according sub-trees (see Fig. 3.).

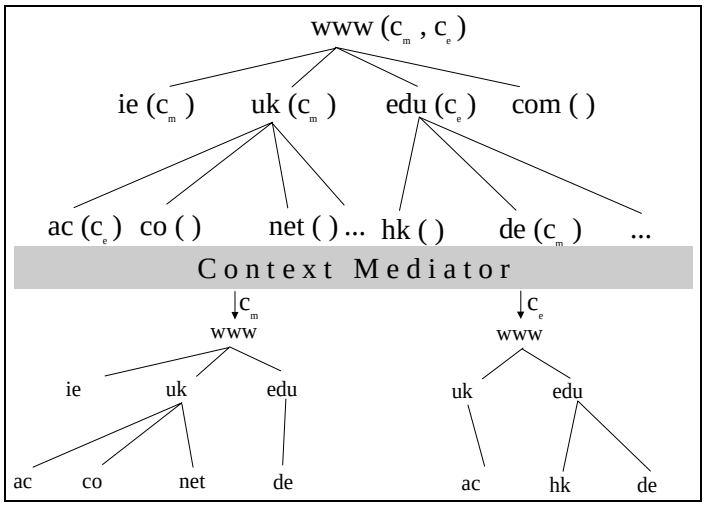

Fig. 3. Example Contextual Concept Hierarchy

Similar to contextual data, contextual domain knowledge has to be mediated. The purpose of the domain knowledge context mediator is to decide what expertise is to be included and what is to be excluded from a data mining task. This decision is based on the context the knowledge has been created in and the context it is to be applied to. These two sites are referred to as knowledge source *s* and receiver *r*. Owing to the fact that a user can only be in one context at a time, the mediator returns the pieces of expertise that have been allotted to the context in which the user is currently in. More formally, this can be expressed as following, where *D* is the set of domain knowledge.

context mediator ß c_r

$$D_r := \mathbf{U} d_s \mid d_s \in D_s \wedge c_s(d_s) = c_r$$

context mediator à D_r

Fig. 4. Domain Knowledge Context Mediator Structure

Contextual Data Mining Architecture

In order to deploy contextual data as well as contextual domain knowledge in data mining applications, a simplified knowledge discovery architecture has been created.

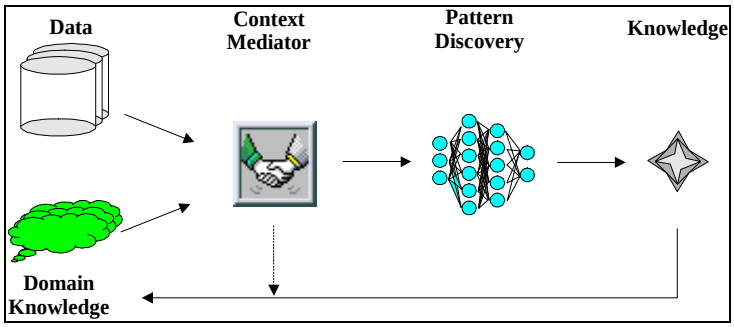

Fig. 5. Simplified Contextual Knowledge Discovery Architecture

The mediator is polymorph, i.e. it can deal with requests about multiple data sources as well as domain knowledge. Depending on the context from which information is requested, data and domain knowledge is used as input for knowledge discovery. The discovered patterns are then contextualised, i.e. labelled with the context of the data mining exercise, and fed back to the domain knowledge repository.

In order to illustrate the operation of the outlined components, consider an e-commerce example, in which sequential patterns (navigational behaviour) are discovered from internet log files. The types of log data as well as marketing-related domain knowledge depend on the type of operating e-tailer. Having site-specific log files (containing data about URLs, login and logoff times, http referrers, statuses, cookies, etc.), the marketing manager as well as the web administrator are looking for interesting patterns. The former has specified her expertise in form of region-based concept hierarchies and target-related age bandings. The latter has created a network topology of the retailer's web site. The thresholds are budget- and cache size-driven, respectively. The discovered sequences are most likely to be different, since they are goal- *and* context-driven. They are kept in the domain knowledge repository after they have been tagged with the current context and memorised for future usage.

References

[1] Sciore, E., Siegel, M., Rosenthal, A.: Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems, ACM Transactions on Database Systems 19 Vol. 2 (1994) 254-290

[2] Büchner, AG., Bell, DA., Hughes, JG.: A Contextualised Object Data Model based on Semantic Values, 11th Int Conf Parallel & Distributed Computing Systems (1998) 171-176

[3] Anand, SS., Büchner, AG.: Decision Support using Data Mining, FT Pitman Publishers (1998)

[4] Büchner, AG., Bell, DA., Hughes, JG.: Contextual Domain Knowledge for Incorporation in Data Mining Systems,. AAAI Workshop Reasoning in Context for AI Applications (1999)