Content-based Retrieval of Analytic Reports

Václav Lín[1], Jan Rauch[1,2], and Vojtěch Svátek[1,2]

[1] Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
{xlinv05,rauch,svatek}@vse.cz

[2] European Centre for Medical Informatics, Statistics and Epidemiology – Cardio
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic

Abstract. Analytic reports are special textual documents containing condensed results from a data mining process. Embedded knowledge enables the interpretation of the reports by automated procedures, which opens the way to content-based retrieval. We elaborate the technique for statistical association rules as specific form of discovered knowledge, demonstrate its formal apparatus on examples from the medical domain, and outline the perspectives of sharing and reusing the content of analytic reports over the Semantic Web.

1 Introduction

The Semantic Web is expected to form a huge, heterogeneous collection of both manually created and (semi-)automatically generated documents. We would like to draw attention to a specific class of documents, which, in a sense, mix the flavors of both types, namely to *analytic reports from KDD* (Knowledge Discovery in Databases). An analytic report from KDD (further only AR) is a textual document presenting the results of a (long and complex) KDD process in a condensed form. The ARs are produced by humans; we also consider the possibility of their automated generation [5]. The core of an AR are the results of data mining procedures; by *result* we also mean the results abstracted, by means of more-or-less deterministic abstraction rules, from the raw output of the procedure. It is clear that ARs are a natural candidate for the Semantic Web. They are well readable by humans, but, due to their regular nature, they can be easily endowed with metadata lending the embedded knowledge (i.e. KDD results) the required machine-processability. The most straightforward application of such metadata is *content-based retrieval* of ARs, much more efficient than keyword-based search.

Throughout the paper, we will use examples from a real-world medical application: results from the analysis of a dataset of 1500 hypertensive patients described by approx. 200 database attributes (columns of the data matrix).

In section 2, we describe the 4ft-Miner procedure we use for the association rule mining, and show an example of AR. The problems related to indexing and retrieving ARs are outlined in sections 3 and 4. Finally, section 5 situates the approach with respect to the related Semantic Web research.

2 Analytic Reports from Association Rule Mining

Association rule [1] is commonly understood as an expression $X \Rightarrow_{c,s} Y$, to read, "X implies Y with confidence c and support s". The procedure 4ft-Miner [7] however mines for *generalized* association rules of the form $X \approx Y$, where $\approx$ – the *4ft-quantifier* – corresponds to a condition on a four-fold contingency table of Boolean attributes X and Y. There are 4ft-quantifiers corresponding to various statistical hypothesis tests (such as χ^2 or F-test) and to various types of equivalence or implication relations between X and Y [6]. *Boolean attributes* are by themselves (usually) complex properties, which are automatically generated from the columns of the input data matrix. An example of such (relatively simple) attribute is the conjunction $Syst\_Blood\_Press(\geq 150) \wedge Cholesterol(too\_high)$.

It is possible to produce several dozens of analytic reports concerning interesting relations among Boolean attributes derived from the database attributes. Each analytic report concerns a general analytic question, or, more often, a group of closely interrelated analytic questions. Examples of those are "*Is there any significant relation between cholesterol levels and skinfold thickness?*" or " *What are the (statistical) implication relations between various combinations of social characteristics and physical activities?*".

Each analytic report informs about interesting implications or equivalences among Boolean attributes; some also contain a part dealing with deviations of the observed attributes' values from their average values. In the examples that follow, we only use relations that are already abstracted from the raw output of the data mining procedure. However, in some cases, the raw output will also be included in the AR. An example of implication relation could be: *"High leisure-time activity (statistically) implies low levels of cholesterol."*. Important part of analytic report will also be "negative" conclusions like: *"No previously unknown implication concerning consequences of high blood pressure was found."*.

ARs have standardized structure with parts *Title*, *Objectives*, *Conclusions* and *Data* being obligatory. They also contain some other parts that concern the data transformations performed, parameter settings, etc. Tab. 1 contains an example of AR. It is a report concerning relationships between biochemical characteristics (i.e. urine, cholesterol, triglycerides) of examined patients. For the sake of brevity, we used only a very simple analytic report, ordinary ARs are much more comprehensive. Furthermore, we give only the very core of the report. Annotations included in Table 1 will be explained in section 3.

3 Annotating Analytic Reports

Since analytic reports describe the results of data mining analyses using statistical association rules, they can be annotated by means of database attributes and classes of statistical association rules. To do so, we introduce hierarchies of *abstract attributes* and *4ft-quantifiers.*

Abstract attributes are representatives of groups of attributes, which follow from the structure of the solved problem. Informally speaking, attributes that are

Table 1. Analytic report - Contents and Annotation

Part	Content	Annotation
Title	Relations among Biochemical Attributes	$\texttt{bio\_atr} \sim_{any} \texttt{bio\_atr}$
Objectives	Are there any significant relations between results of examination of risk biochemical attributes and results of urine examination? Are there any significant relations between cholesterol examination results and triglyceride examination results ?	$\texttt{risk\_bio} \sim_{any} \texttt{urine}$ $\texttt{chlst} \sim_{any} \texttt{trigl}$
Conclusions	No significant relation between results of examination of risk biochemical attributes and urine examination results was found. There is a significant equivalence relation between cholesterol examinations resulting $\geq$ 260 and triglyceride examinations resulting $\geq$ 200.	$\texttt{risk\_bio} \sim_{NOT} \texttt{urine}$ $\texttt{chlst}(\geq 260) \leftrightarrow \texttt{trigl}(\geq 200)$
Data	$\texttt{urine}$, $\texttt{chlst}$, $\texttt{trigl}$	$\{urine, chlst, trigl\}$

being analyzed ”together” form a group which is assigned an abstract attribute. All the original and abstract attributes form a taxonomic hierarchy.

The *4ft-quantifier* $\approx$ relates the Boolean attributes X and Y in association rules $X \approx Y$ the 4ft-Miner mines for. The 4ft-quantifiers are divided into several classes that differ in their logical and statistical properties, see [6]. There are several inclusion relations holding among these classes. We assign an abstract 4ft-quantifier to every class of quantifiers. Inclusion relations among classes allow us to construct a hierarchy of abstract 4ft-quantifiers.

Main parts of ARs usually concern groups of attributes and more-or-less general associations among them. These associations can be expressed by rules of the form $X_A \approx_A Y_A$, where X_A and Y_A are abstract or original attributes, and $\approx_A$ is a possibly abstract 4ft-quantifier.

Tab. 1 shows how a particular AR can be annotated. The abstract attribute $\texttt{risk\_bio}$ represents the group of risk biochemical attributes, $\texttt{risk\_bio} \equiv \{\texttt{trigl}, \texttt{chlst}\}$, while the abstract attribute $\texttt{bio\_atr}$ represents all the biochemical attributes used in the analysis, $\texttt{bio\_atr} \equiv \{\texttt{urine}, \texttt{risk\_bio}\}$.

The symbol $\sim_{any}$ stands for the most general abstract 4ft-quantifier, $\leftrightarrow$ represents the entire class of equivalence 4ft-quantifier and $\sim_{NOT}$ symbolizes the fact that the no interesting association rule was found.

The *Data* part is annotated simply by list of all analyzed attributes. Attributes - results of patients' urine, cholesterol and triglyceride examination are used in Tab. 1. Other parts are annotated with formal sentences composed of (abstract and/or original) attributes and of possibly abstract 4ft -quantifiers, see Tab. 1 We propose that any AR can be annotated this way. Annotation of more complex ARs requires further modeling primitives, but the basic idea of using abstract attributes and abstract 4ft-quantifiers remains the same.

4 Retrieving Analytic Reports

Abstract and/or original attributes and 4ft-quantifiers enable us to retrieve ARs more efficiently than if traditional keywords were used. An elementary query has a form $p(t)$, where p denotes a part of analytic report and t is either an (abstract) attribute or a formal sentence in sense of the previous section. More elaborated queries are formed by joining elementary queries $p(t)$ together by common propositional connectives. In the process of evaluating relevance of an AR to a query, hierarchies of abstract attributes and of abstract 4ft-quantifiers are exploited. We use generalization / specialization relations to decide the relevance. This way, a relevant document can be retrieved even though its annotation does not (syntactically) match the query. Examples of queries and their interpretations are given in Tab. 2. See Tab 1. to compare queries and annotations of the respective AR's parts.

Table 2. Queries - Examples & Interpretations

No.	Query	... denotes a request for:
1	$\mathit{title}(\texttt{urine})$	any AR *concerning* results of urine examination
2	$\mathit{objectives}$ $(\texttt{chlst}(\geq 260) \leftrightarrow \texttt{urine}(\mathit{sugar}))$	any AR which *tries to prove* that patients with sugar in urine or with high levels of cholesterol, tend to suffer from both these symptoms simultaneously
3	$\mathit{conclusions}$ $(\texttt{chlst}(\geq 260) \sim_{any} \texttt{urine}(\mathit{sugar}))$	any AR which actually *proved* some dependency between high levels of cholesterol and presence of sugar in urine

Our example report (Tab. 1) would be evaluated as relevant to query 1, because $\texttt{urine} \in \texttt{bio\_atr}$; any AR that deals with $\texttt{bio\_atr}$ deals also with $\texttt{urine}$. The report is also relevant to query 2, because any AR which aims to analyze problems delimited by $\texttt{risk\_bio} \sim_{any} \texttt{urine}$ (of annotation), will also analyze problems delimited by $\texttt{chlst}(\geq 260) \leftrightarrow \texttt{urine}(\mathit{sugar})$ (of request). To query 3, this report is irrelevant, because it failed to prove the association relation $\texttt{chlst}(\geq 260) \sim_{any} \texttt{urine}(\mathit{sugar})$ requested by the query.

5 Discussion

The techniques of analytic report construction, indexing and (content-based) retrieval, described in the previous sections, are currently tested in the medical domain, in a distributed environment of several clinical and academic sites.

Some of them have previously been applied to other domains such as technical diagnostics. The approach seems to be *conceptually* ready for the integration into the Semantic Web as environment for large-scale knowledge sharing. A hypothetical virtual network of data miners (operating on semantically compatible data), report indexers and query engines could enable synergistic discovery and sharing of interesting relationships empirically valid in the given domain. A necessary prerequisite of the reuse of analytic reports in a non-closed environment is, however, syntactic and semantic interoperability.

The base level of *syntactic* representation is obviously XML. The first, tentative DTD defining the structure of the AR indices (which can possibly be embedded into the textual reports as metadata) has already been designed. The idea of exposing *rule-based* (discovered) knowledge to XML mark-up is obviously not new. The *Predictive Model Markup Language* (PMML, [4]) already includes a sublanguage for 'classical' association rules; it is however not usable for our statistical association rules with automatically generated Boolean attributes, which are more expressive. Another language, *XDMSL* [3], extends the mark–up approach to the whole process of KDD, including the source data model, data transformations, prior domain knowledge and the data mining task description; the structure of *target* knowledge mark–up is again limited to 'classical' association rules, although the inclusion of statistical associations is envisaged by the authors. Finally, a rule mark-up initiative directly associated to the Semantic Web is *RuleML* [2], which offers a suite of languages gradually evolving from purely positional XML to "object-oriented", role-based mark-up (a step towards RDF representation). Our notion of statistical association rule could be added as an extension to the RuleML family of rule types. Furthermore, the problem of bridging the gap between the XML and RDF data models is topical to us: the transformation of our statistical association rules to RDF, necessary for their full inclusion among the Semantic Web knowledge resources, will require several levels of serializations.

Finally, let us mention the *semantic* interoperability, for which *ontologies* are the key enabler. The taxonomies of (original and abstracted) attributes are by themselves trivial ontologies. Richer formalisms would however be needed to facilitate e.g. the integration of ontologies used by the individual clinical sites in our medical application. We consider the use of DAML+OIL [3] for this purpose.

The research has been supported by the project LN00B107 of the Ministry of Education of the Czech Republic.

References

1. Aggraval, R. et al: Fast Discovery of Association Rules. In Fayyad, U. M. et al.: Advances in Knowledge Discovery and Data Mining. AAAI Press , 1996. 307–328
2. Boley, H.: The Rule Markup Language: RDF-XML Data Model, XML Schema Hierarchy, and XSL Transformations. In: 14th Intl. Conf. Applications of Prolog.

[3] `http://www.daml.org/2001/03/daml+oil-index.html`

3. Kotásek, P. - Zendulka, J.: Describing the Data Mining Process with XDMSL. In: ADBIS 2002, Advances in Database Information Systems. Springer Verlag, 2002.
4. PMML 2.0 – Predictive Model Markup Language. http://www.dmg.org/pmmlspecs_v2.
5. Rauch, J.: Logical Calculi for Knowledge Discovery in Databases. In Principles of Data Mining and Knowledge Discovery, Springer-Verlag, 1997.
6. Rauch, J.: Classes of Four-Fold Table Quantifiers. In Principles of Data Mining and Knowledge Discovery, (J. Zytkow, M. Quafafou, eds.), Springer-Verlag, 1998.
7. Rauch, J.: Interesting Association Rules and Multi-relational Association Rules. In Communications of Institute of Information and Computing Machinery, Taiwan. Vol. 5. No. 2, May 2002, pp. 77 - 82