zapthink
zapnote

Doc. ID: ZTZN-0113-1
Released: Dec. 27, 2001

ZAPTHINK ZAPNOTE™

SCIENTIO
MINING XML METADATA

September, 2001

Analyst: Ronald Schmelzer

Abstract

While XML data is fundamentally tree-based and hierarchical, most data that is currently mined is table-based and relational. Many existing data mining algorithms deal don't well with sparse data, and XML is a great example of such sparse data. Scientio has taken an approach to create data mining applications that deal well with the requirements of XML documents.

Mining XML Metadata

Formed in December 2000 by developers working in artificial and computational intelligence, Scientio has taken a fresh approach at data mining, with the requirements of XML square within their sights. Managing the organization, Andy Edmonds has an extensive history of developing applications and performing computational research related to data mining. He fell in love with fuzzy logic a few years ago, and this started him down the path that finally resulted in the suite of solutions presented by the company. Fuzzy logic assists in helping users make decisions for complex processes, such as loan applications.

During this process, one of the things that dawned on him was that XML was a great way to represent information and represent the rules for fuzzy logic systems. Since fuzzy rule induction is very well suited at learning relationships between data elements, it maps into XML very nicely. For example, one can also map neural networks onto XML since XML is a way of representing knowledge, not just data. So, this means that descriptions of rules for logic systems, commonly known as "metarules," can be edited in a straightforward manner rather than just editing the data itself.

Once data is represented in XML, it is possible to "mine" the data to see the inter-relationships and clustering of information. While XML data is fundamentally tree-based and hierarchical, most data that is currently mined is table-based and relational. Many existing data mining algorithms deal don't well with sparse data, and XML is a great example of such sparse data. Sparse data is of the sort where some nodes contain information while others don't. For example, some organizational resource nodes contain salary sub-nodes, while others don't. This provides sparse data that most existing relational data mining systems can't handle.

Scientio has taken an approach to create data mining applications that deal well with the requirements of XML documents. One of their approaches is to use XPath to navigate around XML data in order to mine it. One of the great features of XPath is that it allows for good free-form analysis. While many forms of information can be represented in the tree-based structure of XML, many formats cannot, so it is better to leave that data in their native format. This also allows existing data mining tools to continue to be used along side the Scientio set of structured data mining tools.

The system works by applying mined rules to an XML source file and creating a new copy with the predicted values inserted. Current data mining applications are used to derive data value from massive databases, but the typical use of XML will be in millions of distributed, smaller-scale transactions. So, a system like Scientio will be used to analyze the millions of smaller transactions that happen in eBusiness and similar systems that take advantage of XML. A

typical application would be to evaluate a hundred people that an organization wants to check for credit worthiness. An empty XML element would be created that is populated by the Scientio XMLRule application. XMLRule then populates the field with the predicted value for credit worthiness as well as the degree of belief in that value and a possible range of acceptable values.

Scientio Product Line: XMLMiner and XMLRule

The Scientio product line consists of two major applications: XMLMiner and XMLRule. XMLMiner analyzes and mines XML data. XMLMiner handles the essentially object-oriented nature of XML, in that it copes with missing values, null values and optional values. The system applies Fuzzy Rule induction technology in order to generate a rule set that explains and predicts selected values in an input data set, based on other user selected values. The choice of predicted data values is performed using XPath expressions. The resulting rule set is expressed in METARULE, which is a Scientio-specific XML-based language that can be transformed for other rule-based systems by means of XSL. METARULE rule sets are collections of fuzzy-logic expert system rules, that can also represent Neural nets, symbolic expressions, Boolean logic expressions, Bayesian Belief Networks, and any other form of knowledge representation. METARULE schema can now be found at the XML.org registry.

Applications are also provided to convert METARULE code to plain English if... then rules. Other utilities are supplied to help users create rules and symbolic expressions that can be then used to test XMLMiner's abilities before building it into an application. XMLMiner can read XML strings, entire documents, or URL references. It learns to predict a nominated leaf nodes using fuzzy tree induction followed by rule generation. The result of processing is an XML string conforming to the METARULE schema containing the rule set predicting the nominated leaf node. Input and output nodes can be defined as containing text or index values, or as numeric. A percentage of the data set can be set aside for testing. Users can also specify parent nodes to mine using XPath expressions, leaf nodes containing input data, the predicted leaf node, and can define categorical/numeric status for each input or output or let XMLMiner automatically discover them. Since XMLMiner is a programmatic control that is accessed via COM methods, XMLMiner has no visible user interface.

XMLRule is the runtime processor that interprets and processes the rules created by XMLMiner. The resultant solution can then be embedded in any system that accepts COM objects. Thus, the system produces embeddable code rather than final applications. XMLRule takes the rule sets defined by XMLMiner or separately by users and applies them to any location where a COM control or Java bean can be used. Once a rule set has been loaded, inputs and outputs are available as collections. Input values can be set individually via the collection interface, and outputs read the same way, and confidence values are generated for each output. XMLRule supports logical and arithmetic operators.

Using Mining to Avoid Excessive Hoarding of User Information

One of the trends that is helping adoption of systems such as that produced by Scientio is a political movement to prevent sites from storing customer information. Many existing data mining or other trending systems require extensive profiling of visiting users. Using a general, XML-based data mining approach helps to prevent this hoarding of data by producing generalizations about people rather than specifics. This is the direction many large websites will have to go, and will therefore be a boost to XML-based data mining organizations such as Scientio.

The only real competition for XML-based data mining mechanisms comes from systems that use XML schema for Bayesian belief networks. However, while useful for many artificial intelligence and knowledge systems, Bayesian networks don't do everything that is required

for data mining. In any case, the METARULE system has been created so that it can be converted in and out of Bayesian network systems.

While the Scientio product has yet to be released on a wide scale, they have released their components for download and test by XML developers. They are currently seeking the most appropriate uses for the technology, in a means that allows its components to be embedded in other systems.

Key Conclusions & Recommendations

- Scientio's approach for data mining of XML data is very compelling and should be tested as part of an overall XML data store strategy.
- The Business Intelligence (BI) space is increasingly looking at XML as another market for their tools and technologies, so competition in the space may heat up in the short term.
- Companies with immediate need for XML-based data mining should investigate Scientio tools.

Profile: Scientio (September, 2001)

Date Founded: December 2000
Funding: Privately-held
President: Andy Edmonds
Products:
- XML Miner
- XML Rule
- XR Batch

Address:
Haydon House, Station Road
Woburn Sands, BUCKS, MK17 8RX
United Kingdom

URL: www.metadatamining.com
Main Phone: +(44) 1908-584226
Contacts:
Andy Edmonds andy@metadatamining.com

Related Research

- *XML Data Storage Technologies and Trends* Report (ZTR-ST101)
- *XML Data Storage* Multi-Client Study (ZTR-ST102)
- *Web Services Technologies and Trends* Report (ZT-WEBSRV)
- *B-Bop* ZapNote (ZTZN-0204)
- *Coherity* ZapNote (ZTZN-0144)
- *Excelon* ZapNote (ZTZN-0205)
- *Ipedo* ZapNote (ZTZN-0151)
- *NeoCore* ZapNote (ZTZN-0146)
- *Software AG Tamino* ZapNote (ZTZN-0116)
- *X-Hive* ZapNote (ZTZN-0200)
- *XAware* ZapNote (ZTZN-0154)
- *Xyleme* ZapNote (ZTZN-0326)
- *XYZFind* ZapNote (ZTZN-0117)

About ZapThink, LLC

ZapThink is an IT market intelligence firm that provides trusted advice and critical insight into XML, Web Services, and Service Orientation. We provide our target audience of IT vendors, service providers and end-users a clear roadmap for standards-based, loosely coupled distributed computing – a vision of IT meeting the needs of the agile business.

ZapThink's role is to help companies understand these IT products and services in the context of SOAs and the vision of Service Orientation. ZapThink provides market intelligence to IT vendors who offer XML and Web Services-based products to help them understand their competitive landscape and how to communicate their value proposition to their customers within the context of Service Orientation, and lay out their product roadmaps for the coming wave of Service Orientation. ZapThink also provides implementation intelligence to IT users who are seeking guidance and clarity into how to assemble the available products and services into a coherent roadmap to Service Orientation. Finally, ZapThink provides demand intelligence to IT vendors and service providers who must understand the needs of IT users as they follow the roadmap to Service Orientation.

ZapThink's senior analysts are widely regarded as the "go to analysts" for XML, Web Services, and SOAs by vendors, end-users, and the press. They are in great demand as speakers, and have presented at conferences and industry events around the world. They are among the most quoted industry analysts in the IT industry.

ZapThink was founded in October 2000 and is headquartered in Waltham, Massachusetts. Its customers include Global 1000 firms, public sector organizations around the world, and many emerging businesses. ZapThink Analysts have years of experience in IT as well as research and analysis. Its analysts have previously been with such firms as IDC and ChannelWave, and have sat on the working group committees for standards bodies such as RosettaNet, UDDI, CPExchange, ebXML, EIDX, and CompTIA.

Call, email, or visit the ZapThink Web site to learn more about how ZapThink can help you to better understand how XML and Web Services impact your business or organization.

ZAPTHINK CONTACT:
ZapThink, LLC
11 Willow Street
Suite 200
Waltham, MA 02453
Phone: +1 (781) 207 0203
Fax: +1 (786) 524 3186
info@zapthink.com