zapthink
zapnote

Doc. ID: ZTZN-0121-1
Released: Dec. 27, 2001

ZAPTHINK ZAPNOTE™

AUTONOMY
HELPING TO DERIVE MEANING FOR THE WEB

September, 2001

Analyst: Ronald Schmelzer

Abstract

While a human can identify the similarity between XML tags such as "PO", "PurchaseOrder", and "PurchOrd", an automated system would identify them as differently as "green", "car", and "jump". To address these issues of meaning, or "semantics", Autonomy has developed a set of products that aim to simplify the process of context-aware categorization and tag-set matching. While not a "Semantic Web" technology in the strict sense, it plays in that general environment.

The Need for Better Semantics

Since the emergence of the Internet, the sheer volume of content being exposed and made available to users is increasing exponentially on a daily basis. Of course, the challenge for consumers of that information is how to find and process the most relevant information. Search engines operating on full-text searches return results that are often 90% irrelevant. The problem is even more severe when information is used in business processes. While a human can identify the similarity between XML tags such as "PO", "PurchaseOrder", and "PurchOrd", an automated system would identify them as differently as "green", "car", and "jump". The problem is compounded by differences in context and language. What is French for Purchase Order anyways?

To address these issues of meaning, or "semantics", a number of major efforts have been under way to produce meaning and context-related initiatives. In this space, Autonomy has developed a set of products that aim to simplify the process of context-aware categorization and tag-set matching. While not a "Semantic Web" technology in the strict sense, it plays in that general environment.

Automating Semantics Creation

Autonomy views traditional semantic technology as involving a process that requires a user to build an ontology, derive a taxonomy from the ontology, and then produce a tag set. The main issue in this situation is that human involvement is needed to derive the relationship between terms. The inability to automate this process presents scalability and cost issues.

Autonomy hopes to solve these same issues in a more automated fashion. Ontologies attempt to parse language and the different parts of speech, such as a sentence's subject, object, and verb, and derive its context. However, unstructured text presents unique challenges in trying to the context of what is being described. For example, in the sentence, "The boys drink hot chocolate in winter because its warm", what is being described, boys, hot chocolate, or winter? Similar challenges are faced when attempting to automate systems that depend on the exchange of information. The key to solving this challenge, according to Autonomy, is to look at the patterns, symbolic representations, commonalities, and similarities that are present in any information dialogue.

The Autonomy Content Infrastructure (ACI)

The Autonomy Content Infrastructure (ACI), the company's main product, allows enterprise applications to understand and process the business-critical content that exists in

unstructured formats such as email, Web pages, office documents, and Lotus Notes. One typical usage of ACI is a scenario where a customer sends an email message to a company's technical support complaining about a product defect. The email routing system would first use Autonomy's technology to analyze the message and determine what it is about. The Customer Relationship Management system would then route the email message to the right technical expert for a response, at the same time alerting the salesperson responsible for that account. Meanwhile, the knowledge management system might insert hypertext links to background information on the defect before passing the email message along to the product team responsible for developing the next version of the product. In this manner, no human intervention is required to gain semantic understanding of what the user's requests are about.

The software developed by Autonomy looks at XML as a conceptual entity. Each component, wrapped in XML, is parsed through the engine, which forms a conceptual understanding of text through pattern recognition and statistical matching, rather than using ontology or semantic lookup. Autonomy's strength lies in a unique combination of technologies that employs advanced pattern-matching techniques utilizing Bayesian Inference and Claude Shannon's principles of information theory. Autonomy employs advanced pattern matching technology borrowed from non-linear adaptive digital signal processing to determine the characteristics that give information meaning. Once Autonomy's technology has identified and encoded the unique "signature" of the key concepts, Concept Agents are created to seek out similar ideas in websites, news feeds, email archives and other documents. Because it does not rely on key words, it can work with any language.

Autonomy software identifies the patterns that naturally occur in text, based on the usage and frequency of words or terms that correspond to specific ideas or concepts. The presence of a given pattern frequency enables Autonomy to identify the subject of a given document to a certain degree of confidence. The theoretical underpinnings for Autonomy's approach can be traced back to Thomas Bayes, an 18th century English cleric whose works on mathematical probability were not published until after his death. His work concentrated on calculating the probabilistic relationship between multiple variables and the extent to which they impact one another. Extensions of the theory go further than relevance information for a given query against a text. Adaptive probabilistic concept modeling (APCM) analyzes the correlation between features found in documents relevant to an agent profile, finding new concepts and documents. Concepts important to sets of documents can be determined, allowing new documents to be accurately classified.

In addition, Autonomy makes use of innovations developed by Claude Shannon. His contribution to the space was a discovery that information could be treated as a quantifiable value in communications. The main concept is that most natural languages contain a high degree of redundancy. A conversation in a noisy room can be understood even when some of the words cannot be heard, and the essence of a news article can be obtained by skimming over the text. Autonomy's approach to concept modeling relies on Shannon's theory that the less frequently a unit of communication occurs, the more information it conveys. Therefore ideas which are more rare within the context of a communication, tend to be more indicative of its meaning. Using this theory, Autonomy's software can determine the most informative concepts within a document.

The Dynamic Reasoning Engine (DRE)

The main portion of the ACI system that performs this complex operation is the Dynamic Reasoning Engine (DRE). The DRE performs concept matching, agent creation, retraining, matching, and alerting, categorization, summarization, clustering, active matching, and retrieval, as well as supporting Boolean queries.

The main mode of operation is that Autonomy marks up existing content with XML tags. These tags encode information in a manner so that it can be reused, maintained, and delivered to the proper recipients. Autonomy systems can also be used with potent effect in the supply chain where the system is able to analyze and process related information that may be necessary as part of a purchase order, catalog, or inventory query. Prior to automation of the system, a domain expert will help a user build a taxonomy or ontology based on the quantity and type of data to be automated. The system then crawls through that data, allowing the user to make declarations and use a training set. The system refines itself and concept of topic areas over time. In an XML environment, it then inserts tags automatically in the body of the document. For non-XML systems, the information is made available for further use by portal or other information systems. The software then enables conceptual-based search and retrieval, automated hyperlinking, and a higher level of personalization.

Customers and Release History

Even though the product is expected to be released at the end of the third quarter of 2001, the company has a few early pilot customers who are looking to use the product in an OEM manner. Early adopters include organizations in the financial services and manufacturing industries. There are not as many competitors in this space as in other, more common areas of XML implementation. So far, the major players include mainly new entrants of the like of Applied Semantics, Ontologos, eOntology, and Ontonomy.

Key Conclusions & Recommendations

- The Semantic Web is still an outlying usage of XML that has not yet met with widespread usage. As such, practical implementations as offered by Autonomy can help companies looking to migrate to a semantics-based knowledge representation.
- It is unclear how Semantic Web specifications will develop over time, and as such Autonomy's products may or may not develop in parallel with those efforts.

Profile: Autonomy (September, 2001)

Date Founded: 1996
Funding: Public (LSE: AU.; Nasdaq Europe/Nasdaq: AUTN)
CEO: Michael Lynch, Ph.D.
Products:
- Autonomy Content Infrastructure

Address:
301 Howard Street
22nd Floor
San Francisco, CA 94105
URL: www.autonomy.com
Main Phone: (415) 243 9955
Contacts:
David Appelbaum dma@us.autonomy.com
Linda Findley lfindley@west.schwartz-pr.com

Related Research

- *Epigraph* ZapNote (ZTZN-0139)

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