How taxonomic approaches can alleviate the linguistic problems that arise from universal

Bill Hutchison and Jeremy Ellman
Wordmap Ltd, 26 Upper Borough Walls,
Bath BA2 6LT, United Kingdom.
+44 (0)1225 358182 fax (0)1225 358183
Bill.Hutchison @wordmap.com

Abstract

In describing real world objects and subjects, as well as using the registered names of businesses and products, UDDI will encounter many problems related to linguistics. These include dealing with ambiguous terms, terms with many meanings, synonyms, foreign language variants etc. Approaches which make use of taxonomies can substantially alleviate these difficulties. This paper highlights the major issues and describes solutions.

Keywords: taxonomy, ontology, classification, thesaurus, semantics, ambiguity, synonym.

1. Introduction

UDDI is a proposed standard that describes a methodology for businesses to register, identify and acquire services over the web. Its simple SOAP based approach avoids the fragility and complexity of CORBA and RPC based distribution mechanisms, but the specification has an Achilles heel that should be addressed sooner rather than later. That is, it exploits standard taxonomies to identify goods and services.

Leading users of taxonomic approaches in information management contexts are now realizing that taxonomies can and should be used to process and resolve some of the problems of information retrieval that arise from language.

Collaborative projects such as online marketplaces are dealing with a related set of problems as they try to reconcile differing uses of terms between organizations or communities, large or small.

Common nouns, adjectives and subject descriptors are frequently difficult to map from one organization to another, and there is no ‘dictionary’ on which a given project can standardize. It is likely to be the case that no single standard can be imposed, as cultural differences will in many cases prevail.

Information retrieval taxonomies now incorporate many of the features more usually found in thesauri or dictionaries, such as synonyms and other name variants, foreign language variants, related terms, terms for use in expanded search queries. They also handle a variety of attributes related to each node in the taxonomy. They are cross-referenced to handle ad hoc relations.

The Wordmap taxonomy structure is designed to manage all of the above relationships. The implication of the Wordmap approach for UDDI is to lessen the need for standardization of terminology and increase cross-reference.

This paper proceeds as follows: Firstly, we discuss the problem of word sense ambiguity, and how that applies to product and service specifications for UDDI. Next, we consider what would make a taxonomy adequate for use in UDDI. Finally, we present an overview of the Wordmap taxonomic data structure that addresses these issues. A discussion of future work concludes the paper.

2. Issues

A principal requirement for any user of search services, be that for product or information, is that one can uniquely identify the correct resource. This immediately meets the well-known problem of word sense disambiguation problem (WSD, e.g. Sanderson 1994, Yarowsky 1992). This has two aspects: in the first one word may be applied to several different concepts (which is known as polysemy), and in the second several words may be used for the same concept.

The WSD problem applies equally to product and service descriptions; hence, it is highly relevant to UDDI. Consider the ECMA's UNSPC product classification, which is proposed as a core taxonomy for UDDI. A search here for the vegetable known as "squash" or zucchini in the US (but courgette in the UK), yields two entries: 49161606 Squash Balls, 49161610 Squash Rackets. Both of these clearly refer to the game of squash, and a person can immediately recognize the error. Similarly, staying with the vegetable theme, consider a British English user looking for suppliers of "Ladies Fingers". They are unlikely to connect with Australian okra producers, of Louisiana gumbo farmers, although all are talking about the same product, since none of the three names occurs in the UNSPC.

The situation is little better when it comes to software descriptions and categorizations. NAICS, which is proposed as the industrial classification has three entries here, Programming, Design Services, Facilities Management, and other. Simply, these categories underestimate the complexity of the modern computing industry.

UNSPC, and NAICS are just two of very many categorization schemes. We've used them for the examples above just to illustrate a weak point in the UDDI architecture. We could equally have used the European equivaent categorizations of NACE, and CPV, which are equally general, but have the advantage of being available multi-lingually.

All four classification schemes suffer from the same flaw of a lack of detail. It is each industry requires its own classification scheme. One example here would be EWC –the European Classification of Industrial Waste.

NOTES

RAMON EU Classification Server ¦) Translation of Taxonomies

C. Taxanomic Adequacy

¦) Granularity

NACE has one entry for software consultancy

¦) Specialist Taxonomies

EU Classification Server. Industrial Waste

¦) Proprietary Taxonomies

UNSPC is property of ECCMA ($250) annual fee. NAICS is adapted to Northern American issues

7. Conclusion

There can be little doubt that UDDI is destined for success (Sleeper 2001), as it offers a technically simpler solution to connectivity problems that have been overly technical, fragile, and lack portability. Nevertheless, in the global market place there are potential limitations in the current approach that stem from cultural and linguistic diversity.

References

Bramer M., Macintosh A., and Coenen F (1999). "Research and Development in Intelligent Systems XVI" Springer-Verlag London UK ISBN 1-85233-231-X

Ellman J. 1998 "Using the Generic Document Profile to Cluster Similar Texts." In proceedings 1 St UK Conf. On Computational Linguistics (CLUK).

Ellman J. and Tait T. (1999) "Roget's Thesaurus: An additional knowledge source for Textual CBR?" in Bramer et al. (1999).

Ellman J. (2000) "Using Roget's Thesaurus to Determine the Similarity of Texts" PhD Thesis. School of Computing, Engineering and Technology, University of Sunderland UK.

Green S 1997 "Automatically generating Hypertext by Computing Semantic Similarity" University of Toronto PhD Thesis. Computing Systems Research Group Technical Report 366

Halliday, M.A.K. & Hasan, R.: 1976, "Cohesion in English", Longman, London.

Morris J. and Hirst G. 1991 "Lexical Cohesion computed by thesaural relations as an indicator of the structure of text" Computational Linguistics Vol. 17 (1) pp 21-48

Sanderson M. 1994 "Word Sense Disambiguation and Information Retrieval" in Proc. ACM SIGIR 1994 pp 142-151

Sleeper, B 2001 "Why UDDI Will Succeed, Quietly: Two Factors Push Web Services Forward" The Stencil Group http://www.stencilgroup.com/ideas_scope_200104uddi.pdf

Smeaton, Alan 1999 "Using NLP or NLP Resources for Information Retrieval Tasks" in Strzalkowski 1999

Strzalkowski Tomek 1999 "Natural Language Informational Retrieval" Kluwer Academic, Dordrecht NL. ISBN 0-7923-5685-3

Yarowsky 1992: Word Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. Proc COLING 1992 pp 454-460

Biographical Data

Bill Hutchison is CEO of Wordmap, the UK based specialist in taxonomies. Wordmap technologies were developed in collaboration with leading computational linguist Professor Yorick Wilks.

Dr. Jeremy Ellman is CTO of Wordmap Ltd.