

Ontology-based author profiling of documents

Jan De Bo, Mustafa Jarrar, Ben Majer, Robert Meersman

VUB STARLab
Vrije Universiteit Brussel
Pleinlaan 2
Brussels Belgium
{jdebo, mjarrar, bmajer, meersman}@vub.ac.be
Telephone:+32 2 6293487 Fax:+32 2 6293525
Home Page: <http://www.starlab.vub.ac.be/>

Abstract

In this paper we present the advantages of using an ontology service for the modelling of user profiles in the EC FP5 IST project NAMIC (IST-1999-12392). By means of an ontology server people set up user profiles, which are in fact views, i.e. *specifications of queries* on the ontology. These views are constructed using a JAVA API, which forms the *commitment layer* of the ontology, built on top of an ontology base. In NAMIC an ontology server is used to establish a link between the lexical object representations, generated by the natural language processors (NLP) on the one hand and the user's interest, specified through the selection of relevant concepts and facts of the ontology on the other. This allows to specify a user profile independently of language, categorization and NLP specific "world models". Users then set up a profile consisting of events, agents participating in these events and other content information in which they are interested in. For instance, a journalist writing articles about financial issues may be interested in related documents containing a "raise event" of company shares. If he has specified those conditions in his profile he will be able to retrieve resources which contain events that are semantically related to that kind of event pattern. User profiles in NAMIC do not have to be static. The results of processing by the NLPs of a document the user is currently working on, may be used to construct a dynamic profile, which may contain events specific for that document. This way a user's profile can be dynamically adapted to his current interests. We also developed a tool which illustrates the creation of user profiles using ontological concepts and facts.

1. Introduction and Motivation

In this paper we present results derived from our work in the NAMIC project. Within the NAMIC project the main objective was to develop advanced technologies of Natural Language Processing for multilingual news customization and broadcasting throughout distributed services, which represents one of the major problems for International and National News Agencies (NA) as well as for the spread of Web technologies. Within their own business cases, NAs need to integrate in their own repositories news distributed by other NAs usually in different languages and according to different classification standards. Mismatching is at language level, since different languages are used, as well as at the conceptual, as the organization/storage of news proceeds according to diverging schemes. The volume and richness of this information has, however, a catch: it can overwhelm the pressed user-journalist that may be looking for a particular type(s) of information. This is a well-known problem in an information-rich environment, and especially in the case of (large) sets of hyperlinked documents, often referred to as the "lost-in-hyperspace syndrome".

Several aspects have been researched to improve searching, browsing and retrieval of information. In the information retrieval approach, several techniques ranging from string matching to advanced lexical analyses systems are used in order to understand the implicit semantics and thus the relevancy of the data that will be retrieved. On the other side, in the artificial intelligence and database approaches, such as for example the semantic web, the semantics (and the syntax) of the data are explicitly defined and linked with knowledge bases as ontologies, which help to make precise queries or for reasoning. Experience shows that the accuracy of extracting the

implicit semantics and the relevancy of the data is low, e.g. a search using regular search engines results in a huge amount of information, especially for large volume information resources such as the web, expanding queries to improve recall may also cause huge result sets. On the other hand, defining the semantics of the information explicitly, and reasoning about them in order to retrieve relevant information is an expensive task, and the scalability is very low. Therefore, we believe that combinations of these two approaches will be very fruitful for the improvement of information retrieval, as will be argued in the next sections of this paper.

Within the NAMIC project the User Domain Profiling System (UDPS) allows defining of user profiles for the filtering of news streams according to the specific interests of a user which for NAMIC, primarily would be journalists or other text writers. These user profiles are then used to exclude irrelevant items from a constant stream of documents before these documents are presented to the user.

As will be argued later in this paper, the use of an ontology has critical improvements: IR systems will gain from ontologies richer knowledge representation and modelling capabilities, improved recall by expanding the queries according to well-defined and consistent relationships in the ontology and improved precision by allowing the definition of personalised profile systems as queries against (an) ontologie(s) in order to include or exclude (a) certain type(s) of information.

Structure of the paper. In section 2 we give an introduction of what an ontology is and its critical added value for NLP based systems. Section 3 then gives the definition of a user profile and explains more details about the advantages of using ontology-based information filtering systems such as user profiles. Section 4 demonstrates the implementation done in the Namic Project and Section 5 draws preliminary conclusions and

maps ongoing and future work. Section 6 then places all acknowledgements.

2. Using ontology with NLPs

In this section we will illustrate the advantages of ontologies and their potential role in several aspects of information retrieval and how they can be used in defining user profiles.

Ontology¹ in computer science is a branch of knowledge engineering, where agreed semantics of a certain domain are represented formally in a computer resource, which then enables sharing of information and interoperation between systems. Representing the semantics (as a formal interpretation) of a certain domain implies the conceptualisation of the domain objects and their interrelationships in a declarative way, so that they can be processed, shared, and reused among different applications. Note that an ontology is more than a taxonomy or classification of terms, since it includes richer relationships between terms, e.g. “part-of, location-of, value-of, synonym-of...”(Figure 1). An ontology provides a higher level of knowledge², where the ontology terms are chosen carefully, consistently, and with a higher level of abstraction.

In the DOGMA model described summarily below, we separate relevant ontological relationship knowledge as set extensions of context-specific binary fact types called *lexons*. These express (within this assumed context) plausible relationships between concepts, using lexical terms in a given language; we implicitly assume that these terms are aligned with a lexicon (“terminology base”) that is agreed among all users of the ontology (Jarrar, 2002).

Example. The following –very partial ontology (Tables 1,2,3)- could be lexons in some arbitrary hopefully self-understood syntax, the format for the purpose of textual illustration being (#*contextid*) <term1>[<role label><term2>]; details or omitted in this paper. The ontology base, which contains the set of lexons of the modelled domain, is also known by the symbol, Ω.

(#my_company-ID) employee
is a person
has first_name
has last_name
has empl-id
has birth date
has salary
works_in department

Table 1

(#my_company-ID)salary
is a salary
reviewed_in month

Table 2

(#employment-ID) salary
has amount_in-\$
expressed_in currency
converted_to currency
earned_by employee

Table 3

Through the use of ontologies one is able to express semantic relations between terms, rather than is the case with ordinary categorisations. To express these meaningful relations between different terms we need advanced modelling methodologies, like the ORM conceptual modelling language. We chose ORM for its rich constraint vocabulary and well-defined semantics. Within STARLab we also developed an XML-based ORM markup language (ORM-ML) as a means of exchanging data semantics between different agents. (Demey et al, 2002)

The enormous growth of the Web causes search engines to return a large number of pages to the user for a single search. It is time consuming for the user to traverse the list of pages just to find the relevant information. We claim that information filtering systems based on ontologies will assist the user by filtering the data stream and delivering more *relevant* information to the user. Below are a few examples of how this can be achieved. We will discuss these topics in section 3 in more detail.

IR will benefit from ontologies more than terminology bases/resources since the knowledge is more formally represented than in term bases, which facilitates the representation, maintenance, and dissemination of terminological data and makes these data reusable by computer systems in various applications. Recall and precision of search operations will be improved using ontologies to model the knowledge contained in a system. Recall will be improved by exploiting the rich structure of an ontology and specifying generic queries (Guarino, 1999). The semantics in an ontology makes it quite attractive for query expansion, because there is a strong need to expand queries with relevant terms and meaningful relations which contain a lot of semantics, for instance to include subtopics or to personalize the query according to a user’s personal interests. Precision will be increased through the disambiguation of terms and the ability to navigate through the ontology for the selection of more specific queries (Guarino, 1999).

While ontologies offer highly advanced modelling capabilities our experience indicates that, in the domain of Natural Language Processors (NLPs), ontologies will mostly be lighter, and therefore less expressive, than in other applications such as for example reasoning systems where the reasoning rules (defined as a logical theory in the *commitment layer*; containing for example the following constraint ORM.Mandatory(employee has_birth date)) are the most important part of the ontology, while NLP applications may see the lexons in the ontology base as canonically and linguistically structured expressions.

Furthermore, the context will provide added value to disambiguate (or approximate) the meaning of terms and relations.

Usage of an ontology also offers advantages for multilingual Information Retrieval. Since the ontology is a shared agreement about a (abstract) conceptualization it is in principle independent of a particular natural language

¹ In philosophy, Aristotle defined ontology as the science of being.

² The Knowledge Level is a level of description of the knowledge of an agent that is independent of the symbol-level representation used internally by the agent, (Gruber, 1995)

(assuming the relation between the concepts is SubClassOf).

The ontology is separated from the objective representations used by the natural language processors. Since the user profile is a query on the ontology, this separation hides the user from the potentially large amount of objective representations used by the NLPs. The advantage of the independence between the underlying objective representations and the user setting up his profile is that he does not have to be aware of the different objective representations of the NLPs. The ontology can thus be seen as an intermediate level shielding the different representations of the NLPs from the user. Once the ontology is built, natural language processors will have to adapt their objective representations to it. This way a query on the ontology, can be considered to interact independently with the objective representations generated by various natural language processors.

Because of the multilingual data resources, development of different natural language processors (in NAMIC, English, Spanish and Italian) is required. This was done by the universities of Sheffield, Rome (Tor Vergata) and Catalonia (Universitat Politècnica de Catalunya). The user profiling system, introduced in NAMIC, however enables the user to specify language-independent queries, but still gives the possibility to get back related documents in all languages provided by the news agencies.

As mentioned before a user has the possibility to specify his interests in a static profile by selecting the

appropriate relations and concepts from the ontology. It is however quite possible that a journalist's interests change while working on a particular news story. Therefore the user has to adapt his profile according to his current needs and interests instead of having to create an other additional profile. User profiles, developed within the NAMIC project, can be dynamically adapted. Indeed, as part of the NAMIC profile services, a journalist has the possibility to create a local profile according to the text he is currently working at, because it is likely that he will be interested in retrieving documents containing events, or knowledge related to agents participating in events which he has already entered in his text. The user is given the possibility to update his current static profile according to this new profile, making his own profile change dynamically. This prevents the user from having to manually annotate his own article of text by adding (ontologically derived) concepts and relations to his static profile, assumedly saving time and improving consistency.

4. Implementation

The ontology service in NAMIC provides the possibility to store, edit and retrieve ontological information that models (partial) semantics relevant to the project's domain and in particular the ability to define user profiles based on these semantics.

In order to satisfy the requirements mentioned above we developed a tool, with the following classical two-tier client/server architecture, illustrated in Figure 2.

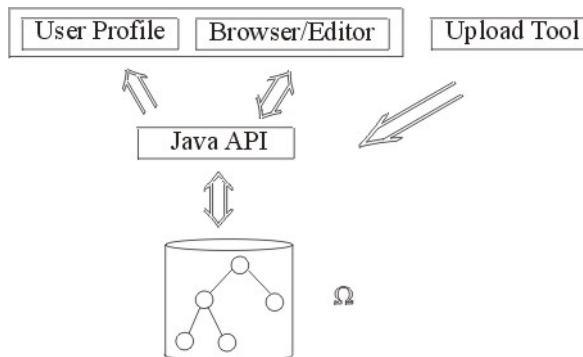


Figure 2

- At the bottom of Figure 2, there is a storage facility for the ontology (in a database)
- Above that, an intermediate API layer establishes communication between various tools and the ontology.
- At the top, support tools like browsers, editors and user profiles are implemented.

In our paper we will use the term 'objective representations' of the natural language processors to refer to Event Matching patterns, which are described in detail in (Basili et al). The process of ontology engineering begins with the development of a base model that provides a framework for the integration of other different, individual resources. The creation of this ontology base can be viewed as a conceptual modelling task, based on ontology merging and alignment of the available resources. The result contains the fundamental concepts based upon the natural language processors'

objective representations, that are generally useful for the project. For instance, consider the following verb syntactic frame: 'person – sells – attribute' as an example of an objective representation from the NLPs' event matching rules. The verb syntactic frame which is not considered to be an ontological concept, is mapped to 'Company Acquisition event'. The occurrence of this verb syntactic frame in a document then results in the detection of a 'Company Acquisition event'.

The individual resources that are considered for their incorporation into the NAMIC ontology were the following:

- The IPTC category system (IPTC)
- The EuroWordNet base concepts (EuroWordNet toplevel concepts) (Vossen, 1998)
- Named Entity lists (Stevenson et al)
- Event Types (Basili et al)

In order to integrate the natural language processors' objective representations of the different individual resources into the ontology, an alignment process needed to be performed between those different representations. Categories, events and named entities are aligned with

EuroWordNet base concepts, by establishing mappings between the involved concepts of the different resources considered for integration in the ontology. This is illustrated in Figure 3; the alignment mappings are depicted as double-sided arrows.

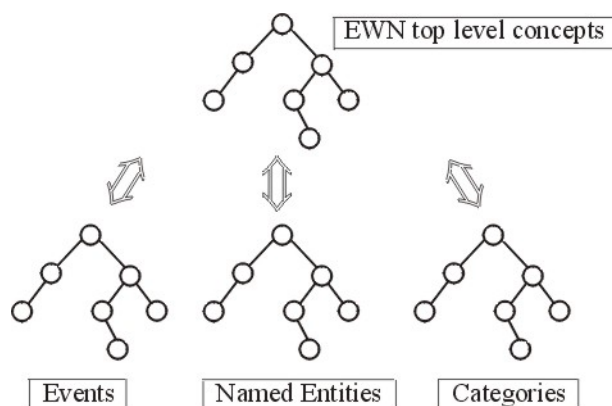


Figure.3

Because an ontology is a *shared* agreement about (a conceptualisation of) the world, aligning different ontologies with one another is required in order to obtain agreement between the concepts of the different ontologies. In order to develop tools automating this activity, good context formalisms will undoubtedly become helpful here but within the scope of NAMIC we had to align the different ontological concepts manually. At this state of the art it is as yet unrealistic to expect that merging or alignment at the semantic level could be performed completely automatically. A prototype of a tool to assist ontology merging and alignment has been built by the Stanford Medical Informatics department of Stanford University. This tool, based on the SMART algorithm, is an extension of the Protégé (Noy, 1999) ontology-development environment.

For the purposes of NAMIC we have also developed a simple custom tool (OntoNAMIC) to make the ontology available for browsing, editing and setting up user profiles.

The browser window consists out of a left pane and a right pane. The left pane is responsible for browsing through the ontology, while the content appearing in the right pane depends on whether one has selected the class view, diagram view or profile view on the toolbar of the application.

When the domain expert (i.e. typically *not* the journalist) selects the Classview, all the lexons containing the selected concept on the left will be displayed in the right pane. Choosing the Diagram view enables one to drag and drop concepts from the left pane into the right. By double-clicking on this dropped concept an ORM diagram appears, displaying all the lexons of which the concept is a part. ORM is a well-known conceptual modelling language (Halpin, 2001) here "re-used" (in part; some interesting modifications are needed that however will be the subject of a separate paper) to represent part of the ontology. In the diagram, ovals represent *entity types*, the rectangles are arbitrary (uninterpreted) relationships

between them, and arrows are (interpreted) *is-a relations*. The important point is that it is possible to map such models to and from lexon-based ontologies, which provides two immediate benefits: a graphical and formally founded notation, and existing tools that already support it, such as Microsoft's VisioModeler for ORM. Because of our earlier experience with this particular method and tools for database design (De Troyer et al, 1995), we have adopted it as a prototypical research and implementation tools and techniques environment for ontology construction.

One then sets up a user profile by choosing the profile view on the toolbar. Remember a user expresses his interests in his profile by specifying a query on the ontology, i.e. as a composition of logical combinations of the desired events, EWN concepts, named entities and categories from the ontology. The resulting implied logical expression will then specify which documents satisfy the profile. This is illustrated in Figure 4.

5. Future work

Although we have now chosen to use a rather simple query language for setting up the user profiles, it is our aim for future work to develop a more sophisticated conceptual query language (for instance similar to RIDL (Verheyen et al,1982)), to specify queries on the ontology.

6. Acknowledgements

This work was supported by the European Commission's IST Project NAMIC (IST-1999-12392). We would also like to acknowledge contributions by our partners in this project Agenzia ANSA S.C.R.A.L., The University of Sheffield, University of Roma Tor Vergata, Universitat Politècnica de Catalunya, Vrije Universiteit Brussel, Comité International des Télécommunications de Presse, Itaca s.r.l., Agenzia EFE, S.A. and Financial Times.

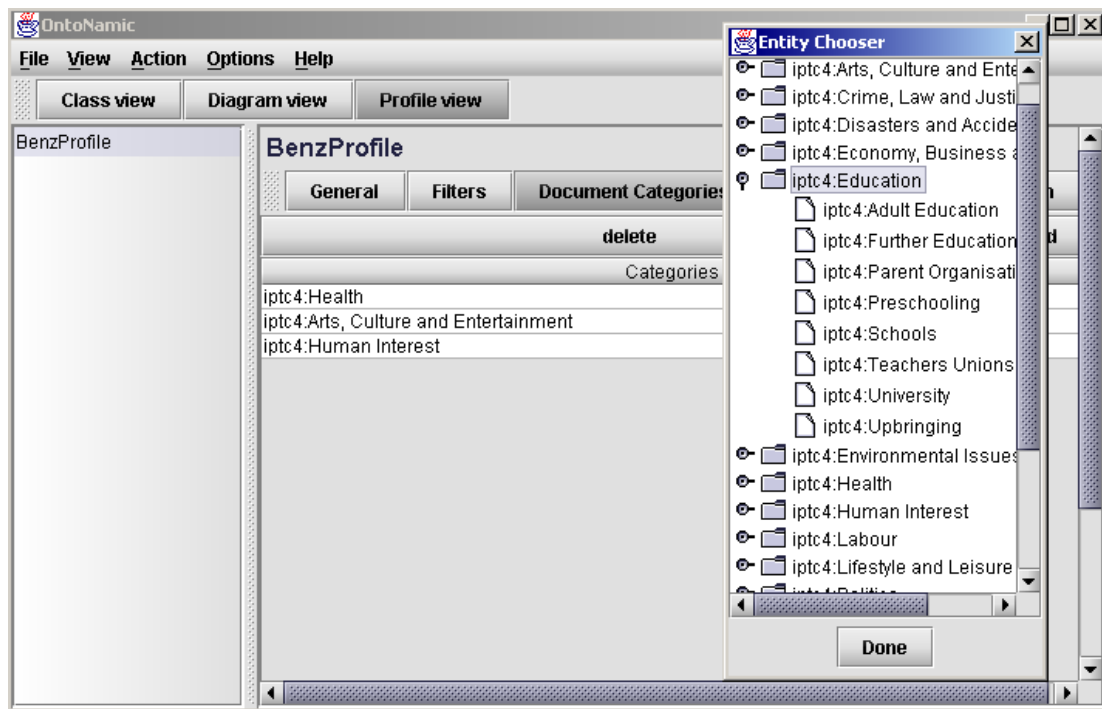


Figure 4

7. References

- Aas K. A survey on Personalised Information Filtering Systems for the World Wide Web; December 1997
- Abuzir Y. and Vandamme F, "E-Newspaper Classification and Distribution Based on user profiles and Thesaurus", SSGRR 2002w - International Conference on Advances in Infrastructure for e-Business, e-Education, e-Science, and e-Medicine on the Internet, January 21 - 27, 2002 L'Aquila (Italy).
- Abuzir Y., Vervenne D., Kaczmariski D. and Vandamme F., "E-mail messages classification and user profiling by the use of semantic thesauri", in CIDE 2001 - in Proceedings CIDE 2001 Conference - 4th International Conference on the Electronic Document, Toulouse - FRANCE Oct. 2001.
- Basili R., Pazienza M.T., Vindigni M. ; Corpus-driven learning of Event Recognition Rules
- Demey J., Jarrar M., Meersman R., Exchanging ORM Schemas Using a conceptual Markup Language, 2002
- De Troyer, O., Meersman, R. : "A logic Framework for a Semantics of Object Oriented Data Modeling" , in: Proceedings of Entity Relationship and OO Modelling Conference, Papazoglou et al. (eds.) Springer LNCS, 1995
- Guarino N., Masolo C., Vetere G., OntoSeek: Content-Based Access to the Web, 1999 IEE
- Gruber T., "Toward principles for the design of ontologies used for knowledge sharing", International Journal of Human-Computer Studies, 43(5/6), 1995.
- IPTC, <http://www.iptc.org/> -> Subjects -> Subject reference system
- Jarrar M., Meersman R., 2002 Practical Ontologies and their Interpretations in Applications – the DOGMA Experiment (to appear)
- Karp P., The design Space of Frame Knowledge Representation Systems, 1993
- Noy Fridman N.. and Musen M., SMART: Automated Support for Ontology Merging and Alignment; 1999
- Stevenson, M. and Gaizauskas, R.; Using Corpus-derived Name Lists for Named Entity Recognition
- Terry Halpin : Information Modeling and Relational Databases: From Conceptual Analysis to Logical Design. Morgan Kaufmann Publishers, 2001. ISBN 1-55860-672-6
- Verheyen, G. and van Bekkum, P., "NIAM, an Information Analysis Method", in: IFIP Conference on Comparative Review of Information Systems Methodologies, T.W. Olle, H. Sol, and A. Verrijn-Stuart (eds.), North-Holland (1982).
- Vossen P (eds), 1998; EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht