Semantically accessing documents using conceptual model descriptions

Terje Brasethvik & Jon Atle Gulla

Department of Computer and Information Science, IDI
Norwegian University of Technology and Science, NTNU
{brase,jag}@idi.ntnu.no

Abstract. . When publishing documents on the web, the user needs to describe and classify her documents for the benefit of later retrieval and use. This paper presents an approach to semantic document classification and retrieval based on Natural Language Processing and Conceptual Modeling. The Referent Model language is in combination with a lexical analysis tool used to define a controlled vocabulary for classifying and indexing documents. Documents are classified using simple sentences …. Classification is done by selecting sentences that contain the highest frequency words in the document that also occurs in the domain model. These are parsed using a DCG-like grammar, mapped into a Referent Model fragment and stored along with the document in RDF-XML syntax. The model fragment represents the connection between the document and the domain model and serves as a document index. The approach is being implemented for a document collection published by the Norwegian Center for Medical Informatics.

1. Introduction

Project groups, communities or even organizations today often turn to the web to distribute and exchange information. While the web makes it fairly easy to make information available, it is substantially more difficult to find a fruitful way to organize, describe, classify and present this information for the benefit of later retrieval and use. In the absence of librarians and library-like tools and mechanisms, it has been up to the users to find a way around this problem.

One of the most challenging tasks is the semantic classification - the representation of document contents. This is usually done using a mixture of text-analysis methods, a carefully defined (or controlled) vocabulary or ontology, as well as a scheme for applying this vocabulary when describing a document. In this paper, we investigate how a given community may perform this manually, aided by a domain model, expressed in a conceptual modeling language and a natural language interface to perform the actual classification and search.

Conceptual modeling languages contain the formal basis that is necessary to define a proper ontology, yet at the same time they offer a visual representation that allows users to take part in the modeling, and to read, explore and use the models. We show that a conceptual modeling language thus represents a vehicle that allows users to define their own domain ontology. Such a domain model then represents a visual definition of terms occurring in the domain and their relations, which at a later stage may be used interactively in an interface to classification, browsing and search. We thus claim that the conceptual modeling language may be used throughout the entire process of presenting documents on the web.

Rather than indexing the documents with sets of unrelated keywords, we then construct indices that represent small fragments of the domain model. These indices do not only tell us which concepts were included in the text, but also how these concepts are linked together in more meaningful units. Natural Language Processing (NLP) is used as the interface to both classification and search. NLP allows users to express clear-text sentences that contain the selected concepts. These sentences are then parsed using a DCG-like grammar and mapped into a model fragment. Searching for a document, the user enters a natural language phrase that is matched against the document indices. Search expressions may be enhanced (enriched/extended, specialized or generalized) through interactions with the domain model.

Approaches to meta-data descriptions of documents on the web today should at least conform to the emerging web-standard for describing networked resources, the Resource Description Framework. We show how our indices of model fragments may be mapped into an RDF representation and stored using the RDF-XML serialization syntax. In our approach, the domain model serves as the vocabulary for defining RDF statements, while NLP represents an interface to creation of such meta-data statements.

The next section of this paper presents the theoretical background behind our approach and references to related work. Sections 3 describes todays situation at KITH[1], the case study that has motivated the approach, while section 4 presents our proposed solution. Our approach is elaborated through a working example from KITH. Section 5 concludes the paper and points to further work.

2. Describing documents on the Web

In many ways, organization and description of information should actually be done by the users. In cooperative settings, users have to take an active part in the creation and maintenance of a common information space supporting their work [1] [2]. The need for - and the use of - information descriptions are often situation dependent. Within a

[1] Norwegian Center for medical Informatics, (Kompetansesenteret for IT i Helse), http://www.kith.no .

group or community, the users themselves know the meaning of the information to be presented, they know the use of the information and are aware of the usage context.

What we should provide, then, is tools to help the users to participate in the organization and description of documents. A definition on how to present information on the web is given in a 3-step reference model by [3]

- Modularization: Find a suitable storage format for the information to be presented.
- Description: Describe the information for the benefit of later retrieval and use. Document-descriptive meta-data may be divided into two categories: Contextual and Semantic descriptive meta-data:
  - *Contextual meta-data:* Covers any contextual property of the document like its author, title, modified date, location, owner, etc. and should also strive to link the document to any related aspect such as projects, people, group, tasks etc.
  - *Semantic meta-data:* Information describing the intellectual content/meaning of the document such as selected keywords, written abstracts/comments and text-indexes. In cooperative settings, or in organizational memory approaches, also free-text descriptions, annotations or collected communication/discussion may be used.
- Reading-functions: On-line presentation of information enables the "publisher" to enable advanced reading-functions such as searching, browsing, automatic routing of documents, notification or awareness of document changes or the ability to comment or annotate documents.

In this paper our focus is the semantic classification of documents using keywords from a controlled vocabulary. What is needed for such an approach, is a way of defining the vocabulary of words - the domain ontology, and a way of applying it when classifying the individual documents. In collaborative settings, what we would like to do is to have users define and use their own domain specific ontology as the controlled vocabulary. Furthermore, what we need is to give users an interface to classification that allows them to work with both the text and the domain model, and to create meaningful indices that connects the text to the model.

When the documents are to be made available to the public, the interface of the retrieval system cannot assume any particular knowledge of the domain. As the users will not necessarily know the particularities of the interface either, the search expressions must be obvious and reflect a natural way of specifying an information need. Adopting natural language interfaces is one way of dealing with this, though the flexibility of natural languages is a huge challenge and can easily hamper the efficiency of the search.

In our approach, we will use a conceptual modeling language – the Referent Model Language – as the users' vehicle for defining and representing the domain ontology. We will turn to Natural Language Processing for the ability to parse selected portions of text, and map these into fragments of the domain model. At a later stage, both the conceptual model and NLP may be used in the user-interface for search, browsing and retrieval of the published documents. These model fragments may then be translated into RDF-XML syntax and published with the document on the web.

2.1 Conceptual modeling and the web – related work

Taking - for a moment - a broad definition of conceptual models into account, we may say that such models are used in several ways for describing and presenting information on the web.

Document type or document structure models and definitions have been used in order to recognize and extract information from text [4]. Data-models of contextual meta-data are used in order to design, set up and configure presentation and exchange on the web [5] [6] [7]. Together with "dataweb" solutions, data models are also used to define the export schema of data that are to populate web-pages created or materialized from underlying databases.

Shared or common information space systems in the area of CSCW, like BSCW [8], ICE [9] or FirstClass [10] mostly use (small) static contextual meta-data schemes, connecting documents to e.g. pepole, taks or projects, and rely on freely selected key-words, or free-text descriptions for the semantic description. TeamWave workplace [11] uses a "concept map" tool, where users may collaboratively define and outline concepts and ideas as a way of structuring the discussion. There is however no explicit way of utilizing this concept graph in the classification of information.

Ontologies [12] [13] [14] are nowadays beeing collaboratively created [15] [16] across the web, and applied to search and classification of documents. Ontobroker [17] [18] or Ontosaurus [19] allows users to search and also annotate HTML documents with "ontological information". Hand-crafted web-directories like Yahoo, or the collaborative Mozilla "OpenDirectory" initiative offers improved searches, by utilizing the directory-hierarchy as a context for search refinement [20]. Domain specific ontologies or thesauri are used to refine and improve search-expressions as in [21]. Domain specific ontologies or schemas are also used together with information extractors or wrappers in order to harvest records of data from data-rich documents such as advertisements, movie reviews, financial summaries. etc. [22] [23] [24].

[25] uses a hierarchy of generic concepts, together with a WordNet thesauri and meta-data in order to organize information and to search in networked multi-databases. Collaboratively created concept-structures and concept-definitions are used in Knowledge Management [26] [27] and Organisational Memory approaches [28]. Many of these approaches use textual descriptions (plus annotations and communications) attached to documents as the semantic classification.

Whereas the early information retrieval systems made use of key-words only, newer systems have experimented with natural language interfaces and linguistic methods for constructing document indices and assessing the user's information needs. The quality of an information retrieval system is often measured in terms of precision (the system does not return documents that you do not want) and recall (the system returns the documents you want). As natural language interfaces use phrases instead of sets of search terms, they generally increase precision at the expense of recall [29] [30]. Some natural language interfaces only take into consideration the proximity of the words in the documents; that is, a document is returned if all the words in the search phrase appear in a certain proximity in the document. In other systems the search

phrases are matched with structured indices reflecting some linguistic or logical theory, e.g. [31] [32]. Finding an appropriate representation of these indices or semantic classifications is central in many IR projects and crucial for the NLP-based IR systems.

Naturally, in order to facilitate information exchange and discovery, also several "web-standards" are approaching. The Dublin Core [33] initiative gives a recommendation of 15 attributes for describing networked resources. W3C's Resource Description Framework, [34] applies a semantic network inspired language that can be used to issue meta-data statements of published documents. RDF-statements may be serialized in XML and stored with the document itself. Pure RDF does however not include a facility for specifying the vocabulary used in the meta-data statements. For this, users may rely on XML namespace techniques in order to define their own sets of names that can be used in the RDF-XML tags. XML [35] are widely used nowadays; for the definition of interchange or export formats from databases, for the definition of document structure and content and in general for storing and exporting meta-data.

3. The KITH example

A Norwegian center of medical informatics - KITH - has the editorial responsibility for creating and publishing ontologies covering various medical domains like physiotherapy, somatic hospitals, psychological healthcare and even the general domain of medical services. These ontologies take the form of a list of selected terms from the domain, their textual definition (possibly with examples) and cross-references among them. Once created, the ontologies are distributed in print to various actors in the domain and are used to improve communication and information exchange in general, and in information systems development projects.

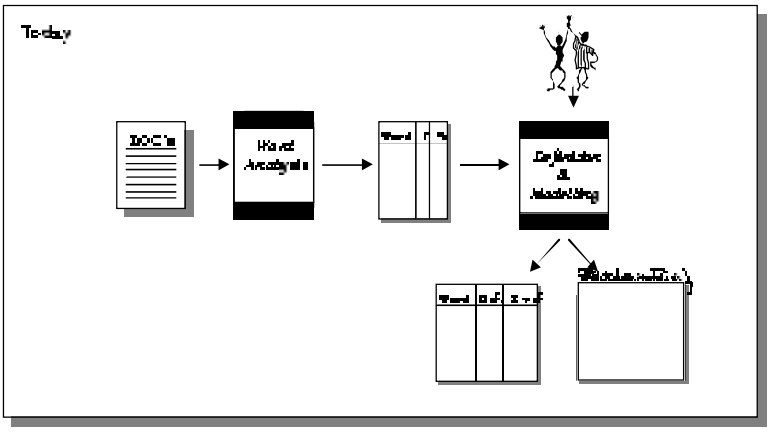

Fig. 1. KITH example – constructing the ontology / modeling process

The ontologies are created on the basis of a selected set of documents from the domain. The process is outlined in figure 1. The set of documents is run through a lexical analysis, using the WordSmith toolkit [36]. The lexical analysis filters out the non-interesting stop words and also matches the documents against a referent set of documents, assumed to contain "average" Norwegian language. The result of the lexical analysis is thus a list of approximately 700 words that occur frequently in (documents from) this domain.

Words are then carefully selected and defined through a manual and cooperative process. This work is performed by a committee run by KITH (computer scientists and medical doctors) that also includes other stakeholders from the particular domain. In some cases, this process is distributed and asynchronous, where participants exchange suggestions and communicate across the web. To some extent, conceptual modelling is used to create an overview of the most central terms and their relations. The models are used as an aid to clarify and structure the set of terms. The final result of this process is a MS Word document containing some overview models and approx. 100-200 well defined terms.

This is where the process stops today. KITH's goal is then to be able to extend this process and to publish and distribute documents on the web, organized and classified according to the developed domain ontology. The challenge is then to exploit the knowledge acquired in the creation of the ontology, and to be able to classify each document accordingly.

4. Referent-Model based classification of documents

An overview of our proposed approach for publishing the medical documents on the web is depicted in figure 2.

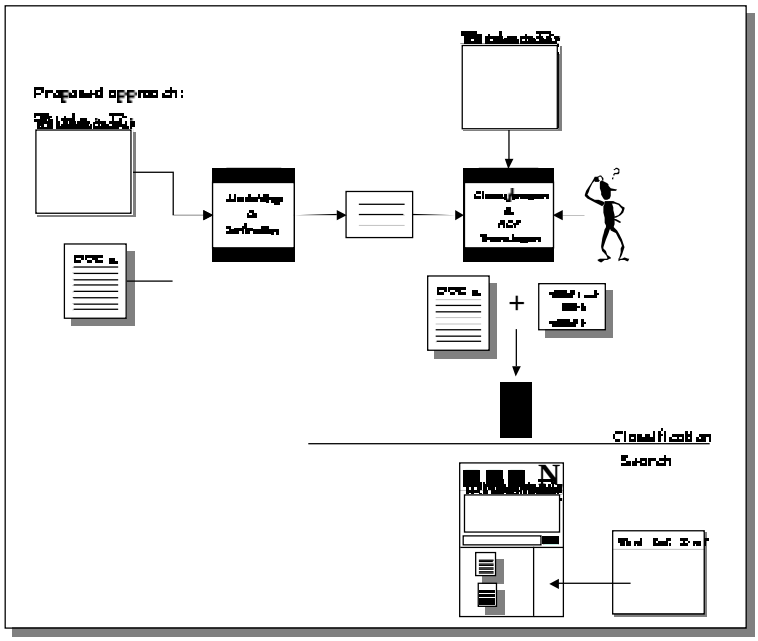

Fig. 2. KITH example: model based classification documents

The classification of documents according to the model is semi-automatically performed following a 2 step process:

1. **Matching:** The document text is analyzed and matched against the conceptual model. The system provides the user with a list of the high-frequency words from the document that are also represented in the model. The user may interact with the model and select those concepts that suits best to represent the document content. The user then formulates a number of simple sentences that contain these concepts. The sentences then relate the concepts in a way that reflects the users impression of the document content.
2. **Classification and Translation:** After formulating the sentences describing the document's content, the user lets the system construct a classification of the document. The system parses the sentences using a Definite Clause Grammar, producing logical formulas for each sentence [37]. These formulas come in the form of predicate structures, e.g. contain(journal, diagnosis) for the sentence "a journal contains a diagnosis", and refer to concepts defined in the domain model. From the formulas, the system proposes possible conceptual model fragments that serve as indices/classifications of the document and are represented in RDF. In case there are several possible interpretations of the sentence in terms of conceptual model fragments, the user is asked to choose the interpretation closest to her information needs. Once the document descriptions and the documents are made available on the web, browsing, navigation and search should be enabled.
3. **Presentation:** Both the model, the definition of terms in the model and - naturally - the documents themselves should be presented on the web. The model represents

an overview of the document domain. Users may browse and explore the model and the definitions in order to get to know the subject domain. Direct interaction with the model should be enabled, allowing users to click and search in the stored documents. Exploiting the capabilities of NLP, also free-text search should be enabled. Search expressions should be parsed and matched against the stored RDF descriptions. In both search approaches, the model may be used in order to refine the search. Search may be specialised or generalised by following subset hierarchies, or users may explore relations and aggregations in order to add more concepts to the search.

4.1 Referent Model Language

Our semantic modelling language - the Referent Model Language - was originally developed at IDI in order to support work in heterogeneous databases and data-warehousing [38]. For these purposes the language was developed with certain features in mind, that we also consider important with respect to our approach:

- The language was given a simple, elegant and compact graphical notation. This is necessary to improve readability of the models. Models must at least be read-able by the end-users, and it is important that they create an overview of terms and their relations.
- The language supports the complete set of CAGA-abstraction mechanisms [39] which is necessary to model some of the semantic relations between terms, e.g. synonyms, homonyms, common & part-terms.

The syntax of the language is shown by example in figure 3. The basic constructs are referent sets and individuals, their properties and relations. These constructs correspond to the need for expressing interpretations in terms of real-world concepts. The set, individual and relational constructs all have their straightforward counterpart in basic set-theory. Properties are connected to referent sets, where each property represents a relation from each member of the set and into a value set.

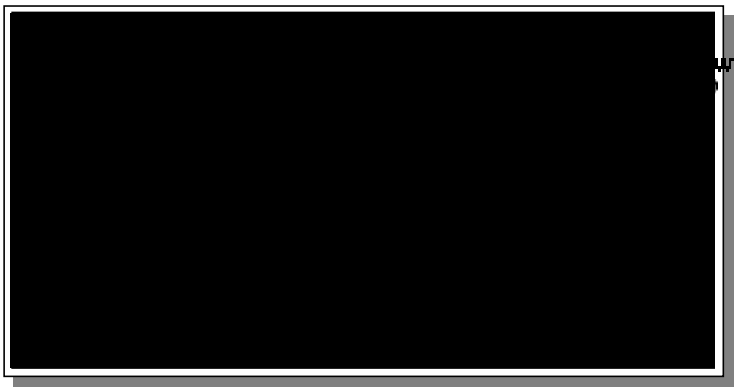

Fig. 3. Referent Model Language, syntax by example.

Relations are represented by simple lines, using arrows to point from many to one and filled circles to indicate full coverage. Several relations may be defined between any two sets. Sometimes relations happens to consist of the same member-tuples, sometimes one relation may be defined as a composition of other relations. Such relations is often referred to as derived relations. Derived relations may be specified using the composition operator o.

The general abstraction mechanisms, Classification, Aggregation, Generalization and Association (CAGA) are all supported by the language. The example to the right in figure 3 shows a world of cats. The set of all Cats may be divided into two disjoint sets male (Catts) or female (Cattes). The filled circle on the disjoint symbol indicates that this is a total partition of the set of cats. That is, every Cat is perceived to be either female or male. On the other hand, there are several other ways of dividing the set of cats, e.g. House cats, Persian cats, Angora cats etc. These sets may be overlapping, i.e. a cat may be both a house cat and an angora cat. The absence of a filled circle indicates that there may be other kinds of cats as well, not indicated in the model (e.g. wild cats, lost cats).

We have also defined a cat as an aggregation of cat parts (heads, bodies, legs and tails). The use of ordinary relations from the aggregation symbol to the different part sets shows how a cat is constructed from these parts. A cat may have up to 4 legs but each leg belongs to only one cat. Heads and bodies participate in a 1:1 correspondance with a cat, that is a cat has only one head and one body. A cat must (filled circle) have a head and a body, while legs and tails are considered optional (no filled circle).

4.2 A working example

Figure 4 shows a test interface for classification. The referent model shown at the top of figure 4(1) is a translated fragment[2] of the model for the domain *somatic hospitals*. The model shows how a patient consults some kind of health-institution, gets her diagnosis, then possibly receives treatment and medication. All relevant information regarding patients and their treatment is recorded in a (medical) patient-journal.

We have then selected a test-document[3], matched it against terms found in the referent model and selected sentences containing words found in the model. The test-applet of figure 4 (2) shows how the user then may have the domain model available (1), work with the sentences (3), select words and fragments of these, and have them translated into RDF statements (4).

---

[2] The term definition catalogue for this domain contains 112 terms and is available in Norwegian only.

[3] *"Regulation for doctors and health institutions' keeping of patient journals"* (in Norwegian), http://www.helsetilsynet.no/regelver/forskrif/f89-277.htm

The translation of NL expressions to RDF starts with a DCG-based parsing of the sentences. The production rules of DCG are extended with logical predicates that combine as the syntactic tree of the sentence is built up. While parsing a sentence like "Journals should be kept for each patient," for example, a logical predicate for the whole sentence is constructed. Some modifications of traditional Definite Clause Grammars are made, so that modal verbs are not included in the logical predicates, and prepositions are added to the verbs rather than to the nominals. For the analysis of the sentence above, thus, the predicate keep_for(journal, patient) is returned. After generating a set of predicates representing the sentences chosen by the user, the system maps the sentences to RDF expressions. Predicates show up as RDF relations, and their arguments are mapped onto RDF nodes. The logical mapping system is based on the abductive system introduced in [40]

This way, the users themselves create RDF-statements that represents the semantic classification of a document. Their own domain model is used as the vocabulary for creating these statements, and these statements are created on the basis of a linguistic analysis of the document.

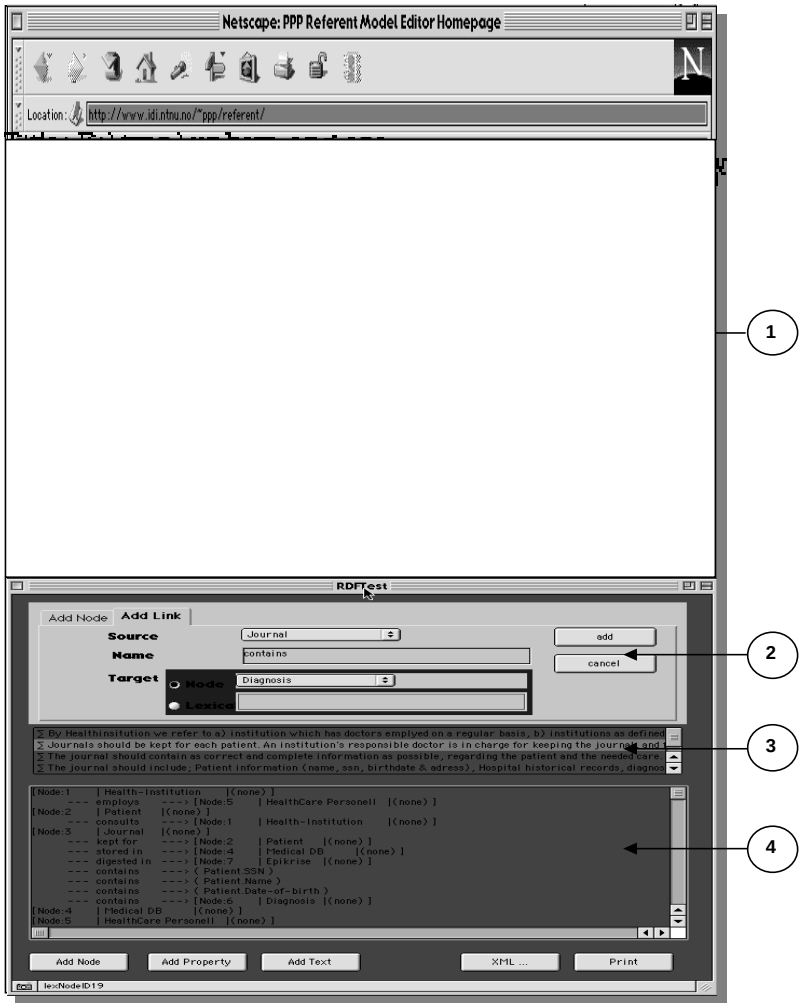

Fig. 4. Selection of sentences and translation to RDF guided by the domain model.

4.3 The RML – RDF connection

The RDF statements represent a connection between the document to be classified and the domain model. The referent model represents the vocabulary for defining RDF-statements. In general, these RDF statements refer to a path in the referent model. Statements are constructed according to the scheme outlined in the sequel.

Words from the selected sentences that have a corresponding referent-set defined in the model are represented as an RDF-node. Also words that can be said to correspond to members of a referent set, may be defined as a node. RDF-properties are then

defined between these RDF-nodes, by finding relations in the referent model that correspond with the text in the selected sentences.

As the original models contain so few named relations[4], we may say that the creation of properties represents an "instantiation" or a "naming" of relations found in the model. In case of naming conflicts between relation names in the model and the text, we use the relation names found in the model. In some cases, we may need to use names on derived relations. This may be done either by selecting the entire composition path as the relation name, or by expanding this path into several RDF statements. The latter implies that we have to include RDF-nodes from referents that may not be originally mentioned in the sentences. Additional information, such as attributes or other "ground" statements is given as RDF lexical nodes. A lexical node need not be mentioned in the referent model.

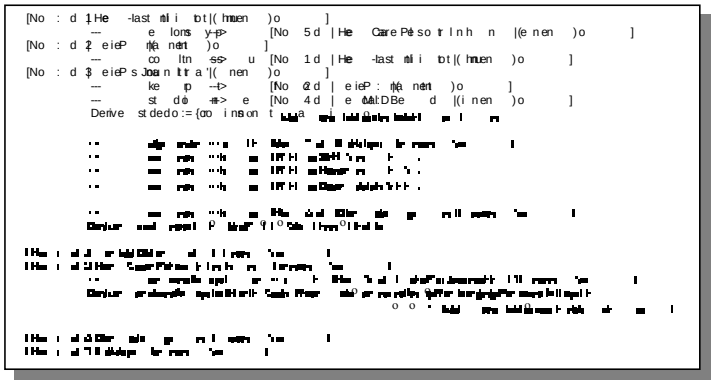

Fig. 5. RDF statements created on the basis of sentences shown in figure 4

Figure 5 shows example RDF statements created on the basis of the model and selected sentences from figure 4(3). The created RDF nodes correspond to the greyed referents in the shown model. In figure 5, derived relations are written, using the composition symbol. The nodes and these derived relations together form a path in the referent model. This path then represents the semantic unit that will be used as the classification of the document. At present, we have not defined mappings for the full RDF syntax. We do not use the RDF-collection constructs, nor the aggregation or meta-constructs. In particular, the aggregation construct should be used to group additional nodes inferred because of derived relations, as mentioned above.

The RDF statements are then translated into RDF-XML serialisation syntax, as defined in [34] (shown in figure 4 (4). Here, the referent model correspondences are given as name-space prefixes to the RDF-XML tags. The XML statements should contain enough references to the original referent model path, to be useful for search and retrieval. In order to export document classifications using these XML

[4] Models in this example are created on the basis of a term catalogue where all relations are denoted "*cross-references*".

representations, we should be able to generate an XML Document Type Definition (DTD) from the referent model path, as well as a XML namespace declaration. This way, these document classifications could be provided for search and retrieval also to external users.

5. Conclusions

We have presented an approach to semantic document classification and search on the web, using a conceptual modeling language and Natural Language Processing. The conceptual modeling language enables the users to define their domain ontology. From the KITH approach today, we have a way of developing the conceptual model that is based on a textual analysis of documents from the domain. The conceptual modeling language offers a visual definition of terms in the domain and their relations that may be used interactively in an interface to both classification and search. The model also provides the formal basis needed of an ontology to be useful for machine-based information retrieval. We have shown through a case example, how the developed domain model may be used directly in an interface to both classification and search. In our approach, the model represents a vehicle for the users throughout the whole process of classifying and presenting documents on the web.

A natural language interface gives the users a natural way of expressing document classifications and search expressions. In our approach, we connect the natural language interface to that of the conceptual model. This way, users are able to write natural language classifications that reflects the semantics of a document in terms of concepts found in the domain. Likewise, users searching for information may browse the conceptual model and formulate domain specific natural language search expressions.

In our approach, the document classifications are parsed by way of a DCG-based grammar, and are mapped into Resource Description Framework (RDF) statements. We may then store these descriptions by means of the proposed RDF-XML serialization syntax. Using RDF, we conform to the emerging web-standard for document descriptions. In our approach, the user developed model represents the vocabulary for creating RDF statements.

Today, we have a visual editor for our Referent Model Language, that may export XML representations of the models. We have experimented with Java-servlets that matches documents against model-fragments and produces HTML links from concepts in the document to their textual definitions, found in the ontology. We also have a Norwegian electronic lexicon, based on the most extensive public dictionary in Norway, that will be used in the linguistic analysis of sentences. Some early work on the DCG parsing has been done, though it has not been adopted to the particular needs of this application yet.

We are currently working on a Java-based implementation. Our goal is to provide a user-interface that may run within a regular web-browser and allows for the classification and presentation of documents across the web. The storage facility for

documents will be provided through a regular web-server, using Java-servlets for the classification and retrieval machinery.

Acknowledgments:

Parts of this work is financed by the Norwegian research council sponsored CAGIS project at IDI, NTNU. Thanks to Babak A. Farshchian for thoughtful reading.

References:

1. Schmidt, K. and L. Bannon, *"Taking CSCW seriously"*. CSCW, 1992. Vol. 1 (No. 1-2): p. 7-40.
2. Bannon, L. and S. Bødker. *"Constructing Common Information Spaces"*. in *5'th European Conference on CSCW*. 1997. Lancaster, UK: Kluwer Academic Publishers.
3. Harmze, F.A.P. and J. Kirzc. *"Form and content in the electronic age"*. in *IEEE-ADL'98 Advances in Digital Libraries*. 1998. St.Barbara, CA, USA: IEEE.
4. Lutz, K., Retzow, Hoernig. *"MAFIA - An active Mail Filter Agent for an Intelligent Document Processing Support"*. in *IFIP*. 1990: North Holland.
5. Hämäläinen, M., A.B. Whitston and S. Vishik, *"Electronic Markets for Learning"*. Communications of the ACM, 1996. Vol. 39 (No. 6, June).
6. Atzeni, P., G. Mecca, P. Merialdo and G. Sindoni. *"A Logical model for meta-data in web bases"*. in *ERCIM Workshop on metadata for Web databases*. 1998. St.Augustin, Bonn, Germany.
7. Poncia, G. and B. Pernici. *"A methodology for the design of distributed web systems"*. in *CAISE*97*. 1997. Barcelona, Spain: Springer Verlag.
8. BSCW, *"Basic Support for Cooperative Work on the WWW"*, http://bscw.gmd.de, (Accessed: May 1999)
9. Farshchian, B.A. *"ICE: An object-oriented toolkit for tailoring collaborative Web--applications"*. in *IFIP WG8.1 Conference on Information Systems in the WWW Environment*. 1998. Beijing, China.
10. FirstClass, *"FirstClass Collaborative Classroom"*, www.schools.softarc.com/, (Accessed: May, 1999)
11. TeamWave, *"TeamWave WorkPlace Overview"*, http://www.teamwave.com, (Accessed: May, 1999)
12. Uschold, M. *"Building Ontologies: Towards a unified methodology"*. in *The 16th annual conference of the British Computer Society Specialist Group on Expert Systems*. 1996. Cambridge (UK).
13. Gruber, T., *"Towards Priciples for the Design of Ontologies used for Knowledge Sharing"*. Human and Computer Studies, 1995. Vol. 43 (No. 5/6): p. 907-928.
14. Guarino, N., *"Ontologies and Knowledge Bases"*, . 1995, IOS Press, Amsterdam.
15. Gruber, T.R., *"Ontolingua - A mechanism to support portable ontologies"*, . 1992, Knowledge Systems Lab, Stanford University.
16. Domingue, J. *"Tadzebao and WebOnto: Discussing, Browsing, and Editing Ontologies on the Web."*. in *11th Banff Knowledge Aquisition for Knowledge-based systems Workshop*. 1998. Banff, Canada.
17. Fensel, D., S. Decker, M. Erdmann and R. Studer. *"Ontobroker: How to make the web intelligent"*. in *11th Banff Knowledge Aquisition for Knowledge-based systems Workshop*. 1998. Banff, Canada.
18. Fensel, D., J. Angele, S. Decker, M. Erdmann and H.-P. Schnurr, *"On2Broker: Improving access to information sources at the WWW"*, http://www.aifb.uni-karlsruhe.de/WBS/www-broker/o2/o2.pdf, (Accessed: May, 1999)
19. Swartout, B., R. Patil, K. Knight and T. Russ. *"Ontosaurus: A tool for browsing and editing ontologies"*. in *9th Banff Knowledge Aquisition for KNowledge-based systems Workshop*. 1996. Banff, Canada.
20. Attardi, G., S. Di Marco, D. Salvi and F. Sebastiani. *"Categorisation by context"*. in *Workshop on Innovative Internet Information Systems, (IIIS-98)*. 1998. PIsa, Italy.
21. OMNI, *"OMNI: Organisaing Medical Networked Information"*, http://www.omni.ac.uk/, (Accessed: May, 1999)
22. Embley, D.W., D.M. Campbell, Y.S. Jiang, Y.-K. Ng and R.D. Smith. *"A conceptual-modeling approach to extracting data from the web."*. in *17th International Conference on Conceptual Modeling (ER'98)*. 1998. Singapore.
23. Gao, X. and L. Sterling. *"Semi-Structured data-extraction from Heterogeneous Sources"*. in *2nd Workshop on Innovative Internet Information Systems,*. 1999. Copenhague, Denmark.
24. Guan, T., M. Liu and L. Saxton. *"Structure based queries over the WWW"*. in *17th International Conference on Conceptual Modeling (ER'98)*. 1998. Singapore.

25. Papazoglou, M.P. and S. Milliner. *"Subject based organisation of the information space in multidatabase networks"*. in *CAISE*98*. 1998. Pisa, Italy.
26. Tschaitschian, B., A. Abecker, J. Hackstein and J. Zakraoui. *"Internet Enabled Corporate Knowledge Sharing and Utilization"*. in *2nd Int. workshop IIIS-99*. 1999. Copenhague, DK: Post workshop proceedings to be published by IGP.
27. Voss, A., K. Nakata, M. Juhnke and T. Schardt. *"Collaborative information management using concepts"*. in *2nd International Workshop IIIS-99*. 1999. Copenhague, DK: Postproceedings to be published by IGP.
28. Schwartz, D.G., *"Shared semantics and the use of organizational memories for e-mail communications"*. Internet Research, 1998. Vol. 8 (No. 5).
29. Stralkowski, T., F. Lin and J. Perez-Carballo. *"Natural Language Information Retrieval TREC-6 Report"*. in *6th Text Retrieval Conference, TREC-6*. 1997. Gaithersburg, November, 1997.
30. Schneiderman, B., D. Byrd and W. Bruce Croft, *"Clarifying Search: A User-Interface Framework for Text Searches"*. D-Lib Magazine, 1997. Vol. (No. January 1997).
31. Katz, B. *"From Sentence Processing to Information Access on the World Wide Web"*. in *AAAI Spring Symposium on Natural Language Processing for the World Wide Web*. 1997. Stanford University, Stanford CA.
32. Rau, L.F., *"Knowledge organization and access in a conceptual information system"*. Information Processing and Management, 1987. Vol. 21 (No. 4): p. 269-283.
33. Weibel, S. and E. Millner, *"The Dublin Core Metadata Element Set home page"*, http://purl.oclc.org/dc/, (Accessed: May 1999)
34. W3CRDF, *"Resource Description Framework - Working Draft"*, http://www.w3.org/Metadata/RDF/,
35. W3CXML, *"Extensible Markup Language"*, http://www.w3.org/XML/, (Accessed: May 1999)
36. Scott, M., *"WordSmith Tools"*, http://www.liv.ac.uk/~ms2928/wordsmit.htm, (Accessed: Jan 1998)
37. Pereira, F. and D. Warren, *"DCG for language analysis - a survey of the formalism and a comparison with Augmented Transition Networks"*. Artificial Intelligence, 1980. Vol. 13 (No. : p. 231-278).
38. Sølvberg, A. *"Data and what they refer to"*. in *Conceptual modeling: Historical perspectives and future trends*. 1998. In conjunction with 16th Int. Conf. on Conceptual modeling, Los Angeles, CA, USA.
39. Hull, R. and R. King, *"Semantic Database Modeling; Survey, Applications and Research Issues"*. ACM Computing Surveys, 1986. Vol. 19 (No. (3) Sept.).
40. Gulla, J.A., B. van der Vos and U. Thiel, *"An abductive, linguistic approach to model retrieval"*. Data and Knowledge Engineering, 1997. Vol. 23: p. 17-31.