The ALVIS Document Model for a Semantic Search Engine

Wray L. Buntine and Kimmo Valtonen
Helsinki Institute for Information Tech. (HIIT)
P.O. Box 9800, FIN-02015 HUT, Finland
First.Last@Hiit.FI

Michael P. Taylor
Index Data Aps.
Købmagergade 43, 2, 1150 København k.,
Denmark
mike@indexdata.com

ABSTRACT

ALVIS is a research project in the design, use and interoperability of topic-specific search engines with the goal of developing an open source prototype of a peer-to-peer, semantic-based search engine. Existing search engines provide poor foundations for semantic-based web operations. Our approach is not the traditional Semantic Web approach with coded metadata, but rather an engine that can build on content through semi-automatic analysis. This paper describes the ALVIS document processing architecture that supports the processing of documents for these purposes.

1. THE ALVIS OBJECTIVE AND RATIONALE

The ALVIS objectives are: to provide a powerful, free, stand-alone semantic-based search system so that application-domain experts can readily build topic-specific search sites without needing to become information retrieval experts or computer systems gurus; and also to also develop complementary distributed components, together with bridges to existing topic-specific search sites, so that the individual sites can be linked up to form a search network. The semantic-based search engine is intended to automatically build and maintain its own semantic structure with named entities, topics and so forth, and to input primitive ontologies. It is not a Semantic Web engine, and does not rely on the existence of Semantic Web ontologies or build its own ontologies. The semantic structure is created semi-automatically using statistical and machine learning methods for the purpose of returning better search results. The distributed system is intended to be able to operate with heterogeneous search servers, using query topics as a routing mechanism, and using distributed methods for ranking and semantic-based processing.

There are four main factors that motivate this design [2]: The individual search engines we provide *(1) must have more capabilities than existing major commercial search engines.* A user must gain significant advantage from ALVIS services over standard search. ALVIS will make subject specific search sophisticated enough to motivate interest groups. If ALVIS can *(2) harness the efforts and imagination of talented groups and individuals in the research and development community*, then the effort can be maintained after the end of the funded project's lifespan. Open Source effort is essential to match the significant resources of the commercial engines. The so-called deep web provides a large, rich set of resources across many areas both commercially and in digital libraries. The *(3) deep web is naturally accessed by a distributed system*, and not by a single centralized indexer. Finally, in the light of the distributed open source development of Linux, and the Open Directory Project, *(4) we view open source as a healthy business model for this particular service*, important in terms of encouraging participation.

2. STANDARDS FOR OPEN SOURCE PROCESSING OF DOCUMENTS

The basis of ALVIS, then, is the individual peer, capable of taking in documents, analysing them semantically and building suitable indexes, then making those documents available across the wider ALVIS network. Within each peer, a processing pipeline generates the semantic information. Document processing is then summarised in the architecture of Figure 1. Routine document processing occurs before hand over to the run-time search system, and subsequently the peer-to-peer system. Collection-wide analysis and maintenance of the system's semantic and linguistic re-

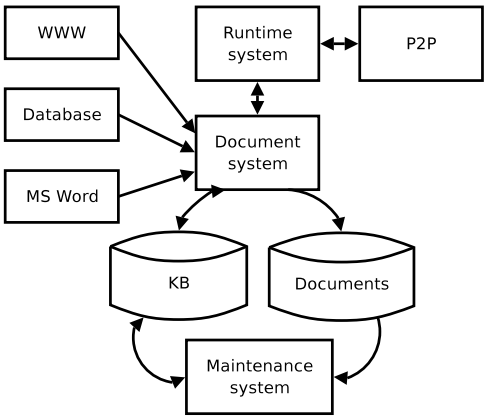

Figure 1: Overall architecture

ESWC '05 Heraklion, Crete

sources then occurs in on off-line maintenance system, that uses the document representation.

It is in the maintenance system that tools for semi-automated ontology learning would be used, for instance, we are starting to experiment with tools for simple ISA-hierarchy discovery [4]. Note we currently restrict our semantic links to the ISA type and use RDF as the standard for representation.

We are currently in the process of readying and documenting the XML formats used in the pipeline so that we can integrate a wide variety of linguistic and annotation tools, and allow others to do so as well.

2.1 Document Acquisition and Canonicalisation

Documents may in principle be of any type and acquired from any source: HTML documents harvested by a Web crawler, local PDF files scanned from the local disk, MS-Word files from a company-wide repository, etc. In order that they can be handled uniformly by the later started in the pipeline, document acquisition models are required to generate a canonicalised version of each document, conforming to a simple XML schema. The details of how this canonical version is generated vary depending on the source-document's format. The simple scheme only recognises sections and lists, eliminating other information at present. Conversion from HTML to this form is simpler than full application of W3C's `tidy` for conversion to XHTML, known to fail on a significant percentage of documents. Thus HTML such as:

```
<HTML><BODY>
<TABLE cellpadding="20" cellspacing="20">
<TR><TD>
Tell me, Alvis!<BR>
You're the dwarf who knows everything ...
</TD>
<TD> ...
```

is converted to:

```
<canonicalDocument>
    <section>Tell me, Alvis!
    You're the dwarf who knows everything ...
    </section>
    <section>...
```

2.2 Linguistic Processing

Linguistic processing within an ALVIS peer is done on the canonicalised version of the documents, and produces as its result a more complex XML document that includes both the canonical document itself and a series of stand-off annotations [1] in a format borrowing from ISO proposition (TC37SC4/TEI). Full natural language processing is notoriously slow, but information retrieval systems make use of partial methods, for instance shallow parsing and tagging. We expect different subject areas to require different levels of linguistic processing: for example, an ALVIS peer specialised for zoological data might include facilities for recognising and tagging formal biological names.

Standoff annotation records data separately to the original text, so that alternative versions can co-exist in the one document. Thus "Cloudless Swill" in the original becomes

```
<c_alpha><cont>Cloudless</cont>
    <from>100</from><id>token1</id><to>108</to>
</c_alpha>
<c_sep><cont> </cont>
    <from>108</from><id>token2</id><to>108</to>
</c_sep>
<c_alpha><cont>Swill</cont>
    <from>109</from><id>token3</id><to>113</to>
</c_alpha>
```

and named entity recognition results in

```
<semantic_unit>
    <named_entity><id>sem_unit1</id>
        <form>Cloudless Swill</form>
    </named_entity>
</semantic_unit>
```

This is course needs to be matched with good compression (eliminating significant redundancy) to keep the XML documents in a reasonable size.

2.3 Relevance

Relevance is the technology used in ranking documents for a given query. The ALVIS relevance engines understand and augment the output of the Linguistic Processing step. Relevance calculations includes static ranking of documents (the best known example is Google's PageRank), as well as automatic categorization of documents for topic-based retrieval. Topics are document-level semantic annotations (rather than term-level annotation such as a named entity). We use discrete PCA [3] to develop multifaceted topics that are then named and placed in a hierarchy by hand. This is significantly enhanced by using the linguistic outputs so that topics can be characterised by nouns, verbs and noun phrases[1].

3. CONCLUSION

ALVIS is intended to bridge the gap between the unstructured web and the Semantic Web, and between existing topic specific search engines. We have developed an open framework for processing documents in ALVIS that allows a variety of linguistic and annotation tools to be integrated into a document processing pipeline that produces XML for our semantic-based search engine. The framework above was tested by running it on 1.5Gb of text from the Wikipedia using a lemmatizer, and discrete PCA to develop a 100 node two-level topic hierarchy. The resultant search engine is be prepared for public use.

Acknowledgements

This overview is from internal reports written by partners in the consortium. ALVIS is funded as contract number 002068 in Information Society Technologies from 6th Framework Programme.

4. REFERENCES

[1] E. Alphonse, S. Aubin, J. Derivière, T. Hamon, D. Mladenic, A. Nazarenko, C. Nédellec, T. Poibeau, D. Weissenbacher, and Q. Zhou. Report on method and language for the production of augmented document representations. ALVIS Deliverable D5.1, ALVIS, 2004.

[2] W. Buntine. Open source search: A data mining platform. *SIGIR Forum*, 39, 2005. To appear.

[3] W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *UAI-2004*, Banff, 2004.

[4] C. Nédellec. Ontologies and information extraction. In S. Staab and R. Studer, editors, *Handbook on Ontologies in Information Systems*. Springer Verlag, 2004.

[1]Demonstration at `http://cosco.hiit.fi/search/MPCA`