ALVIS: SUPERPEER SEMANTIC SEARCH ENGINE

WRAY L. BUNTINE

Complex Systems Computation Group (CoSCo),
Helsinki Institute for Information Technology (HIIT)
University of Helsinki & Helsinki University of Technology
P.O. Box 9800, FIN-02015 HUT, Finland
E-mail: wray.buntine@hiit.fi

MICHAEL P. TAYLOR

Index Data Aps.
Købmagergade 43, 2.
1150 København k., Denmark
E-mail: 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, and are becoming monopolies, distorting the entire information landscape. 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.

1 Introduction

The Alvis Consortium is building a Superpeer Semantic Search Engine as a Specific Targeted Research Project in the EU Commission's 6th Framework Programme. Our view is that search is a fundamental service, and thus should become part of the public domain infrastructure that is the web. We also view search and information access in the context of the Web as a grand challenge problem, one that rivals the great challenges of physics in its potential impact to society. By opening up a distributed search infrastructure built on semi-automatic technology for handling semantics, the technical community is likely to spawn a flood of new capabilities and new services for the public. Moreover, in the light of the distributed open source development of Linux, and the Open Directory Project, we view open source as a healthy business model for this particular service. It is within this vision that we have undertaken the Superpeer Semantic Search Engine project.

2 The Alvis Objective

The Internet holds many individual collections of high quality. These collections and their search engines are specialised and optimised in their handling of languages, subject domains and ontologies as well as document formats.

Our starting point, then, is twofold: (1) 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; (2) 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 resultant network might have components as shown in the Figure 1. 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, and is not intended to be an explicit representation of knowledge.

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.

3 Outline and Rationale

The Alvis design relies on three critical factors:

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*

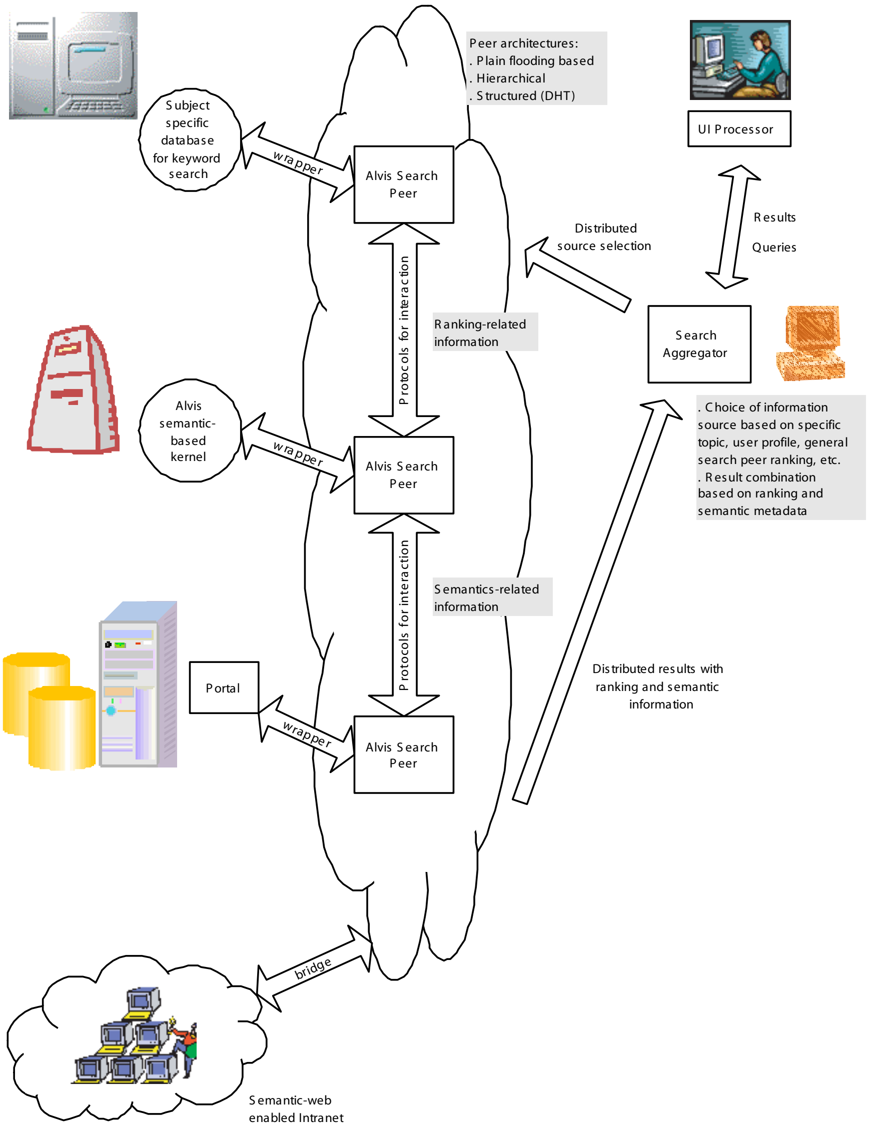
Peer architectures:
. Plain flooding based
. Hierarchical
. Structured (DHT)
UI Processor
Subject specific database for keyword search
wrapper
Alvis Search Peer
Results
Queries
Distributed source selection
Protocols for interaction
Ranking-related information
Search Aggregator
. Choice of information source based on specific topic, user profile, general search peer ranking, etc.
. Result combination based on ranking and semantic metadata
Alvis semantic-based kernel
wrapper
Alvis Search Peer
Protocols for interaction
Semantics-related information
Portal
wrapper
Alvis Search Peer
Distributed results with ranking and semantic information
bridge
Semantic-web enabled Intranet

Figure 1: Distributed Architecture

naturally accessed by a distributed system, and not by a single centralized indexer.

The Alvis approach is to treat each individual semantic-based search engine as a peer in a peer-to-peer network, and route queries to the peers that will be best able to respond to them, using an adaptive approach to continually refine the routing strategy.

To help meet its objectives, the Alvis consortium includes expertise in information retrieval, digital libraries, peer-to-peer systems, information extraction, and probabilistic modelling from the following organizations: Helsinki University of Technology, Helsinki Institute for Information Technology HIIT, Institut National de la Recherche Agronomique,Unité Mathématique, Informatique et Génome, Ecole Polytechnique Fédérale de Lausanne, Distributed Information Systems Lab, Lund University, Department of Information Technology, Technical University of Denmark, Center of Knowledge Technology, Index Data Aps, Exalead SA, University Paris 13, Laboratoire d'Informatique de Paris-Nord, ALMA Bioinformatics, S.L., Jozef Stefan Institute, Department of Intelligent Systems, Tsinghua University, Department of Computer Science and Technology.

4 Standards for Open Source Processing of Web 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:

4.1 *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.

4.2 *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. 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.

4.3 *Relevance*

Relevance is the technology used in ranking documents for a given query. The Alvis relevance engines understand and work with the annotated XML format produced by the Linguistic Processing step. They produce documents in yet a third XML format, which includes the canonical document, the linguistic annotation and the results of the relevance calculation. 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.

4.4 *Indexing Engine*

The enriched documents are indexed ready for subsequent retrieval; as a part of the retrieval process, relevance-enabled plug-ins are invoked to generate suitable "snippets" for result-set summaries. The Alvis architecture does not mandate the use of any specific indexing engine, but the prototype system will use the open source Zebra engine (http://www.indexdata.com/zebra).

4.5 *User Interface*

We have developed an XML format for representing search results. This has the ability to incorporate the grouping of documents, auxiliary annotations, and auxiliary categorizations found in many current avant garde search engines. For instance, keyword information, topic categorization, geographic information, relevant names of people or organizations, or other auxiliary information could be included in this format.

5 Current Work

Initial Alvis effort has focused on several areas:

- Developing several candidate distributed (peer-to-peer) architectures, focusing on the coupled indexing and retrieval tasks and on query routing.

- Preparing document collections, large and small, specialized and general, for early testing. For instance, a one terabyte collection representing the top level web (according to the Open Directory Project) is being converted to the Alvis canonical format described above. The Wikipedia, and a collection about Mosquitos form other test sets.

- Developing protocols and interfaces for component systems described in the previous section.
- Components are being integrated for named-entity identification: people, places, companies, etc.; and for keyword identification: keywords are phrases special to a particular topic area that can be identified at a corpus level.
- Relevance is the assessment of results. Documents retrieved are scored according to some measures intending to evaluate their relevance to the user and their specific query. A system for relevance is being developed based on the above linguistic processing as well as the topic framework.

6 Conclusion

Alvis is intended to bridge the gap between the unstructured web and the Semantic Web, and between existing topic specific search engines. By joining topic-specific search engines - especially those specialised for related but distinct disciplines such as zoology, botany and paleontology - it will enable synergetic use of resources that were previously isolated from each other, so that the whole of an Alvis peer network will be greater than the sum of the parts.

Acknowledgements

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