Overview

ALVIS is a European project building a semantic-based peer-to-peer search engine. A consortium of eleven partners from six different European Community countries plus Switzerland and China contribute expertise in a broad range of specialities including network topologies, routing algorithms, probabilistic approaches to information retrieval, linguistic analysis and bioinformatics.

The consortium members are the coordinator Helsinki Institute for Information Technology; a number of internationally reknown institutes and university laboratories: 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; Université Paris-Nord, Laboratoire d'Informatique; Jozef Stefan Institute, Department of Intelligent Systems and Department of Knowledge Technologies; and Tsinghua University, Department of Computer Science and Technology; and a vibrant mix of small business: Index Data Aps.; Exalead SA; and ALMA Bioinformatica, S.L

See the table below for the key facts about the project.

<table>
<thead>
<tr>
<th>Title of contract</th>
<th>ALVIS - Superpeer Semantic Search Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acronym</td>
<td>ALVIS</td>
</tr>
<tr>
<td>Contract number</td>
<td>IST-1-002068-STP</td>
</tr>
<tr>
<td>Start date of the project</td>
<td>1.1.2004</td>
</tr>
<tr>
<td>Duration</td>
<td>36 months, until 31.12.2006</td>
</tr>
<tr>
<td>Countries</td>
<td>Denmark, Finland, France, Slovenia, Spain, Sweden + China, Switzerland</td>
</tr>
<tr>
<td>Project websites:</td>
<td><a href="http://www.alvis.info">www.alvis.info</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://www.opensourcesearch.org">www.opensourcesearch.org</a></td>
</tr>
</tbody>
</table>

Objectives

The ALVIS project is developing a search engine as open source. Since global search engines require vast scale and resources, the project has instead chosen to focus on a modest self-sustaining development path. To this end, two complementary activities are being pursued and integrated: open tools for building high-quality subject specific search engines, and an infrastructure for supporting a peer to peer network. The subject specific engines will provide some semantic awareness using information extraction technology, to make the system attractive for our initial targeted user base: organizations, academic groups and digital libraries. The peer-to-peer infrastructure will allow inter-operability, the sharing of search resources, and the distributed solution of search tasks.

Thus, the objectives for the ALVIS project are:

1. A subject area specialist in Europe could run a small topic-specific search engine with superior quality over general search engines.

2. The technology will have been developed so that should enough of these engines be installed, they can be tied together for general access through a distributed protocol so that users can obtain search services without having to know the individual sites.

3. Adequate scientific studies are made to provide some assurance that the system so developed can indeed operate effectively, and scale adequately since we cannot build a full European-wide distributed-system ourselves.
Summary of Activities

The project, now in its second year, had its annual meeting in January 2005 in Paris, and a working architecture meeting was held in beautiful Bohinj in Slovenia in June. The ALVIS consortium has established its initial code base, test plans, and some public demonstrations, although is not quite ready with a configurable distribution and publicly accessible demonstration web sites.

Resolving a number of architectural issues, developing initial code, and preparing test plans was the core effort for the first half of the year. The resultant architectural plan and the XML standards used for communicating document data and metadata are on the architecture web-page\(^1\).

The second part of the year was concerned with furthering prototypes, preparing for testing in 2006, and preparing basic demonstrations and documentation for a number of successful dissemination events held in Autumn. For instance, the Wikipedia demonstrator has proven itself a useful conceptual tool. The first round of test preparations and code refinement was also done, although the major part of this effort is to follow. A basic software release site is the ALVIS software page\(^2\), and individual components will be maintained separately by the individual authors.

Details of Main Activities

The overall architecture of the ALVIS system is given in Figure 1. The major components are:

**Input system:** a variety of source formats - HTML, DocBook, PDF, MS-Word, etc. - have their salient content converted into a simple canonical XML format. Dynamic aspects (“onclick”, etc. in HTML), comments and scripts etc., unusable by text search engines, are removed.

**Document system:** a domain-specific linguistic analyzer annotates the canonical document with the results of segmentation and lemmatisation processes, and adds notes of domain-specific technical terms, named entities, and identified topical content that it has recognized. This enriched document is further augmented by a relevance engine, which adds domain-specific scores, and specifies content for indexing.

**Maintenance system:** collection-wide analysis to develop linguistic and semantic resources. This supports the document processing activities of the document system.

---

\(^1\) [http://www.alvis.info/alvis/architecture](http://www.alvis.info/alvis/architecture)

\(^2\) [http://www.alvis.info/alvis/software](http://www.alvis.info/alvis/software)
Superpeer: this is a stand-alone runtime for search for indexing and query handling at the peer, and standards for the user interface.

Peer-to-peer system: distributed indexing and query handling.

Documents in the system use an XML format defined with an XML schema, where more or less entries are present depending on the stage of processing. This means ALVIS is an open architecture; components to the system can be added if they can work in the right XML formats, using for instance XSL style-sheets to perform data transformations.

Indexable documents are thus added to the repository and enter the run-time and peer to peer systems. The run-time system has its own set of standards:

- The Common Query Language (CQL) serves as a simple query language with good open source support in the way of software tools.
- The Search Retrieval via URL (SRU) standard carries CQL queries and provides a URL standard for interacting with search engines, together with a simple XML standard for packaging results.
- An extended results format has been adopted since the emerging standard, OpenSearch RSS, lacks tagging for critical result metadata. This is to be embedded in SRU results.

The architecture was implement in 2005 in a number of demonstrations, but a fully configurable packaged system is still under development.

The peer to peer infra-structure: The P2P infrastructure is based on a novel retrieval model that builds a full-text global key-to-document index in structured P2P overlays. We index keys—rare and discriminative terms and term sets appearing in a limited number of collection documents. The size of posting lists is restricted which directly addresses the main problem of P2P information retrieval, namely, unscalable bandwidth consumption.

We have developed a set of distributed algorithms for building the key index and performing search in a P2P environment. The algorithms rely on self-organizing properties of P2P overlays and enable each peer to index a local document collection independently with respect to other peers. We have defined an open and layered system architecture with a Web Service interface allowing other systems to use the P2P engine for distributed indexing and search. The prototype implementation and experimental evaluation of our approach was the major effort performed.

Subject specific search engines: While designing the pipeline and architecture above, we paid special care to allow a support infrastructure for the development of these subject specific search engines. In 2005 we have been further developing techniques to support this, including the adaptation of parsers to specific domains, the development of topic hierarchies to support organization of content, the analysis of genre/style of a web page, and the easy bootstrapping of subject specific crawlers.

Testing and demonstrators: The linguistic processing developed by Paris 13 and INRA will be tested on some MEDLINE content, to explore the impact of, for instance, named entity extraction at the interface. A previous subject specific search engine, Materials.DK, will act as the subject for longitudinal tests. A search engine about search engines will be used to exercise a number of interface and search techniques using the results of natural language processing. Finally, the P2P system will be tested on a range content for both computational performance and accuracy.

User interface activities: We implemented the first running prototype of the user interface, and a demonstration was made in September 2005. The user interface is a key enabler for our focus on subject specific search and poor-man’s semantic web. We built a generic infrastructure for dynamic data and text mining on Alvis search results, which can be used to propose new kinds of user interactions.
Packaging and release The task of code packaging and release documentation was begun. The indexing system, focused crawler and a basic ALVIS pipeline has been packaged as a Debian package and the same process is now being repeated for other components, such as linguistic processing. Release of code for a fat-peer is thus beginning.

Promotion

ALVIS members had a busy year promoting the project at a number of major European and international events. More details of specific events and papers are available at the ALVIS publications web-page.


It has also been a busy year for open source search in general with the 2005 Workshop on Open Source Web Information Retrieval, held prior to WI 2005. ALVIS members were present at this lively event and we all look forward to the next one in 2006 we have proposed coincident with SIGIR 2006. At the following 2005 IEEE/WIC/ACM International Conference on Web Intelligence, Compiègne, France, on 21 September 2005, an invited talk was given on “Opportunities from Open Source Search” by Wray Buntine. This distilled some of the discussion from the previous open source workshop.

ALVIS is combining with some members of SEKT (Semantically Enabled Knowledge Technologies) and the PASCAL Network (Pattern Analysis, Statistical Modelling and Computational Learning) to co-host a number of events in 2006 to support our shared interests, including workshops in ontology learning and information access. One workshop, Intelligent Information Access\(^3\), will be our major dissemination event for the project.

For more information:

See our project websites:

- http://www.alvis.info :: main ALVIS site.
- http://www.opensourcesearch.org :: theme site

Or contact the administrative or technical representatives, both at HIIT:

<table>
<thead>
<tr>
<th>Administrative Contact</th>
<th>Mr Mikko Kontiainen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tel:</td>
<td>+358 9 451 8157</td>
</tr>
<tr>
<td>Email:</td>
<td><a href="mailto:Mikko.Kontiainen@hiit.fi">Mikko.Kontiainen@hiit.fi</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technical Contact</th>
<th>Dr Wray Buntine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email:</td>
<td><a href="mailto:Wray_Buntine@hiit.fi">Wray_Buntine@hiit.fi</a></td>
</tr>
</tbody>
</table>

Fax: +358 9 694 9768

Mail address: HIIT, Dept. of Computer Science, PL 68, 00014, University of Helsinki, Finland

\(^3\)http://cosco.hiit.fi/search/IIIA2006/