Star Software Corporation

Partnerships Built on Agile Solutions

Star Software: Best Practices

Client: Millennium Pharmaceuticals, Inc.
Project title: MedWizard
Project domain: Information Retrieval System in Molecular Biology
Business domain: Large-Scale Information Retrieval

Background

Millennium Pharmaceuticals (Millennium) is a large American pharmaceutical company specialized in bio-molecular research and new pharmacological agent discovery. The success of bioinformatics partly depends on the availability of computer-readable information assisting and guiding any experimentation.

The Medline database is one of the most authoritative collections of abstracts of medical publications. However, apart from string pattern search in these textual data, there is virtually no way of making their contents computationally available for analysis or retrieval.

Most approaches to the retrieval of electronically available material depend on a lexical match between words in users request and those in database objects, which is a major barrier to successful retrieval from external sources, as different authors can describe the same objects in different ways.

In summer 1999 Millennium expressed their interest in continuing co-operation with Star that resulted in setting terms for a new project to develop an Information Retrieval (IR) system to process MedLine. It was agreed that the SWS Thesaurus (developed in course of Star-Millennium first project) should be used as main terminological resource.

Project scope

This project is limited to the domain or research of proteins and genes. Therefore, from the entire collection of MedLine abstracts only those were processed which belong to this domain.

This project is based on terminological resources crafted in previous projects, which Star has completed for Millennium. In addition, a set of tools will be provided to allow semi-automatic extraction of terminology from the text corpus and interactive editing of the Thesaurus.

The system IR engine is built on a concept-based mining technique. Terms in the source text are morphologically normalized and then looked up in a Thesaurus of concepts (sets of synonyms). If found, they are substituted with their parent concept. Each abstract is indexed with all concepts found in it. Certain other factors are taken in account in these indices. For example, frequency of concepts in a given abstract and in the entire text corpus, calculated weight of concepts, and so on.

The scope of the current project includes:

(1) Development of the updated Thesaurus. Extension of the Thesaurus with terminology extracted from the text corpus.
(2) Creation of the MedW index entry for each abstract in the target corpus selected as relevant to the MedW subject domain.
(3) Development of a search engine using the MedW index and providing term- and concept-oriented query facilities.
(4) Development of a user interface (Medline Wizard) providing an access to search and reporting functions.

(5) Development of the Thesaurus Editor (TEd) based on the Star's automated terminology extraction technology (Star-TEx).
(6) Development of the MedW Product Quality Plan (PQP) in order to test the software and to benchmark the search engine's recall and precision.
(7) Feasibility study for the next phase of the MedW software development: conceptual graph approach and Information Extraction perspective.

Product requirements

Streamline access to the knowledge

- Automatic processing of huge volumes of textual data using extended NLP (Natural Language Processing) techniques.
- Sophisticated Web-based information retrieval system characterized by complex query language and semantic document similarity metrics.
- Using standard http-access, system should be easily integrated into the existing resource- and knowledge management frameworks.
- Links to particular documents are optionally integrated into user's pages.

Information classification

- Information classification facilities extract information on all concepts that co-occur in the resulting document set and report this information to the user in a convenient form

Open architecture

- The module architecture of the system allows concurrently improving terminological resources and continuing the main retrieval process.
- Open architecture is friendly accept new plugs-in, e.g. new terminological resources, new processing modules (e.g. acronym processing module)

Newsagent

- Newsagent registers user queries and when new updates to the MedLine are available it performs searches over new abstracts and sends results (if any) to the user that has issued the query.

Users profiles

- The system remembers previous queries, thus minimizing the time required to generate report (queries are stored as browser's cookies).
- User can set preferences (e.g. particular journal or magazine) for each query or remember preferable sets in user profile.

Technologies

Environment: Sun Solaris, IRIX, Linux, Windows 95/98/NT
Technologies and Languages: Client/Server, CORBA, Web-agents, Servlets, Java, JSP, C++, DHTML, JavaScript
Tools: JDK 1.2.2

Project methodology

The project started in November 1999. Star developed the Functional Specification document approved by Millennium. Star responsibilities included all the project life-cycle activities: system design, coding, testing and user documentation development.

There were certain challenges that put the project into a class of its own among Star projects:

Huge volumes of data to be processed. The MedLine database volume is 24 Gb – pre-processed to filter out the protein/gene relevant part of it. 6 Gb of protein-related MedLine abstract texts were processed and indexed.

Multidisciplinary project team. The project tasks implementation required that a team of biologists, computational linguists, mathematicians, system analysts and programmers should work together contributing to a common task.

New intricate field of knowledge. The project has been the first Star experience in development a system for biomedical texts processing that are characterized by a overcomplicated multi-word terminology and complex syntax.

Sophisticated NLP techniques application. In course of the project implementation a wide range of Natural Language Processing techniques were used: tokenizing, morphological normalization, pattern matching, etc

Project summary

Duration	phase I - 8 months; phase II – 10 month
Team size	phase I – 10 persons; phase II – 6 persons
Efforts	140 person/month
Pricing	Combined Fixed Price and Time & Materials

About Star Software

Star Software is a leading Russian software-outsourcing provider specializing in the implementation and maintenance of information systems. On November 15, 2002, CIO Magazine named Star Software among the Top Three Offshore Software Developers in Russia. See www.cio.com/offshoremap/russia.html

Star offers particular expertise in database-intensive applications, migration of legacy systems to web-based environments, application maintenance and software localization. For corporate Knowledge Management solutions, Star Software offers proprietary data mining tools based on NLP (Natural Language Processing) techniques.

Former and current clients of Star Software include CSC/Denmark, IBM/Tivoli, Millennium Pharmaceuticals, Contex Scanning Technologies, STAC, Tupperware, Foss Electrics, LISA (Localization Industry Standards Association), and UNU (United Nations University). Among the end-customers for the developed software are Berghof Muhlhausen, Hugo Boss, Adidas, Schreyer, Danish Ministry of Labor, Danish Ministry of Tax and Customs, and many others.

Star Software: Contact information

Star Software Corporation

Address: P.O.Box 70, St. Petersburg 197101, Russia
Phone: +7-812-327 9900
Fax: +7-812-327 9865
E-mail: info@Star-sw.com
Web: http://www.Star-sw.com