
Swift Search Software and Services (4S)

Javad K. HESHMATI / Rudy DEMO ♦ 23-3-2006 ♦ 10 Pages



Dixite

Av. Louise 179, P.B.: 3, 1050 Brussels ♦ www.dixite.com

Summary

Presenting Swift Search Software and Services (4S) which is a full-featured text search engine written entirely in Java 2 Platform, Standard Edition (J2SE).

CONTENTS

0.1. Introduction	3
0.2. Index Engine	3
0.2.1. Description	3
0.2.2. Features	3
0.2.3. Architecture	4
0.3. Search Client	5
0.3.1. Description	5
0.3.2. Features	5
0.3.3. Architecture	5
0.4. Case Study - CURIA II	7
0.4.1. Project Name	7
0.4.2. Project description	7
0.4.3. Technical description	7
0.4.4. Project References	8
0.5. Acronyms	10

About Dixite

Dixite is an Internet-based Engineering company established since 1999.

Dixite develops and maintains state-of-the-art systems and has a profound experience with private and public corporate customers. Dixite is also committed, whenever possible, to use open source software to satisfy its customers's requirements.

Since 2000, Dixite has been active on the market of *Index & Search* software and services. Check our professional references on www.dixite.com.

0.1 Introduction

4S is the result of our expertise on *Index and Search Services*.

4S is a full-featured text search engine written entirely in J2SE. It provides text indexing and searching capabilities which makes it suitable for nearly any application that requires full-text search, especially multi-lingual and cross-platform.

The scope of this document is to provide an overview of 4S main features and briefly describe how it functions.

0.2 Index Engine

0.2.1 Description

The *Indexation Engine* is a text indexer. Indexing is the process of creating the *index*. The index is a special database containing a compiled version of terms and documents pointers and is optimized for fast documents access via keywords lookup. Indexation is performed based on a set of files in various formats organized as shown in figure 0.2.3 on the following page.

0.2.2 Features

A non-exhaustive list of *4S Index Engine* features is:

- Multiple and expandable document formats support. It currently supports HTML, PDF, Text, RTF and XML;
- Fields analysis support: a document to index is a set of fields handled separately. Field examples are: document content, author, keywords, etc. and each of these can be handled separately;
- Field-based manipulation: allows the combination of fields indexing, storage and tokenising;
- Incremental indexation;
- Document-based index administration ;
- Stop-words support;
- Multilingual support: via Unicode native support and any character set;
- Stemming;
- multi-platform support; Including support for Mac, Windows NT/2K/XP, Unix and Linux.

0.2.3 Architecture

As shown in figure 0.2.3, the *4S Index Engine* uses *4S Analyser* to read documents in various formats and languages.

- 4S provides various content feeders which may be adapted to correspond to your specific content;
- The *Document Repository* itself can be created on proprietary and non-proprietary file formats;
- 4S *Analysers*¹ recognise the file formats and extract the document content to be indexed;
- 4S *Indexer*² creates the indexes;
- The *List of Indexes* is a set of files compiling terms and document pointers. Various index files are necessary in order to support various and powerful searches capabilities.

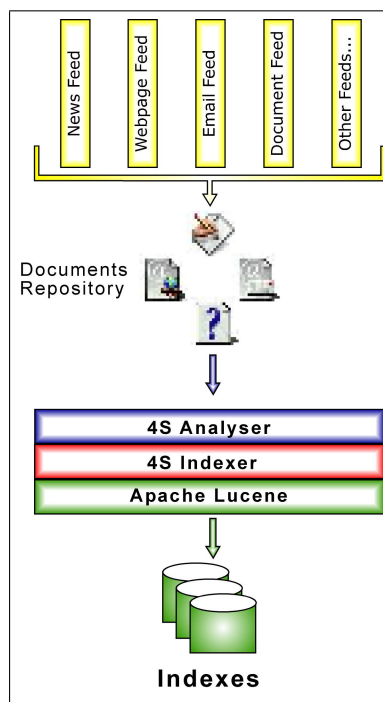


Figure 1: Index Process

¹Currently, 4S uses standard Lucene analysers but customised analysers may be provided.

²4S uses Apache Lucene indexation engine.

0.3 Search Client

0.3.1 Description

Searching is the operation of locating a subset of documents which contain an expressed content. The input for a search operation is a *query* specifying some criteria. Once the query is executed, the result set is shown back along with navigation capabilities.

0.3.2 Features

A non-exhaustive list of *4S Search Client* features is:

- Term or Phrase query support;
- Field search support: allows searches within parts of the document. For example, document content or its properties like author, keywords, etc.;
- Term modifiers: through wildcard usage;
- Boolean operators support;
- Queries grouping;
- Fuzzy, proximity and range values search: Fuzzy search allows one to search for a term accepting certain approximations. A proximity query finds document whose term is enclosed within a certain term distance. Range query hits results within which values are enclosed between bounds;
- Results ranking: allows the presentation of each result relevance;
- Term boosting: modifies the relevance of a term within a query expression;

0.3.3 Architecture

Depending on the requirements several technologies may be used to query the *index* produced by the *4S Index Engine*. See figure 0.3.3 on the next page

- *4S Search engine* can be used to query off-line or on-line document repositories. By default, the standard query syntax is used. However, this query syntax may be adapted to your requirements.
- *4S* can be a stand-alone application or be integrated within various desktop applications³. *The 4S Java Applet* is the search client⁴ provided by default.
- An *Analyser* is used before submitting the query to the index, in order to ensure that the same transformations have been applied to the indexed documents and the query.

³like mail client, web browser, version and content management, news service, site grabber, etc.

⁴can be used from any Java enabled web browser

- The search client functionalities can be developed using various *programming languages*⁵.
- The *indexes* contain statistics on terms and document pointers in order to lead to performant document searches.

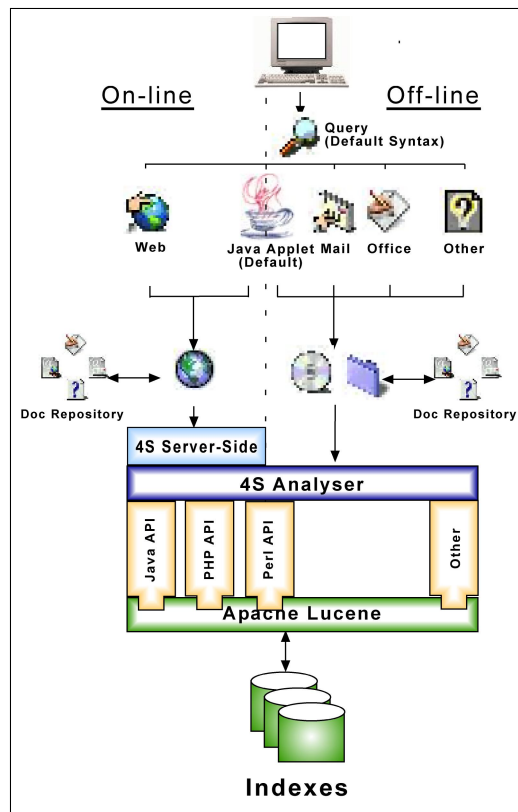


Figure 2: Search and Query Process

⁵Java, Hypertext Preprocessor (PHP) or other Common Gateway Interface (CGI) based applications maybe used to perform queries.

0.4 Case Study - CURIA II

0.4.1 Project Name

CURIA II - Office des Publications de la Communaut Européenne(OPOCE) and Court of Justice

0.4.2 Project description

Dixite has been awarded the development of an application that would allow the monthly incremental publication of the case-laws of the European Court of Justice and European Court of Instance on a CD-ROM media.

Based on the CURIA site of the European Court of Justice, similar document management functionalities have been implemented but on a CD-ROM media. The CD-ROM is generated for each of the 11th languages of the European Union. The data are incrementally added over the 12 monthly edition of the CD-ROM.

CURIA II is a high-performance, fully-featured search engine written in Java. It includes features as:

- Boolean and phrase queries;
- Field searching;
- Ranking and sorting results;

The architecture is based on *open-source* Java libraries.

0.4.3 Technical description

Methodological Approach

Object Oriented Analysis and Design, XML based technologies, CD-ROM based technologies, Document Management Systems.

Material

GNU/Linux, Windows NT4/2000 and Mac OS X computers.

Software

- Java
- XML/SAX
- XSL/XSLT/DTD
- as libraries:

- Lucene (Apache Group)
- Xerces (Apache Group)
- JAXP (SUN)
- Castor (Exolab)

0.4.4 Project References

<http://curia.eu.int/jurisp/cgi-bin/form.pl?lang=en>

Document Information

Copyright

This document can be freely redistributed according to the terms of the GNU Free Documentation License (GFDL). To learn more about GFDL, visit the following *URL* ↗ www.gnu.org/copyleft/fdl.html.

0.5 Acronyms

CGI Common Gateway Interface

HTML HyperText Markup Language is the lingua franca for publishing on the World Wide Web. Having gone through several stages of evolution, today's HTML has a wide range of features reflecting the needs of a very diverse and international community wishing to make information available on the Web.

J2EE Java 2 Platform, Enterprise Edition

J2SE Java 2 Platform, Standard Edition

JSP Java Server Pages

PHP Hypertext Preprocessor is a widely-used general-purpose scripting language that is especially suited for Web development and can be embedded into HTML.

PDF Portable Document FormatA file format developed by Adobe Systems. PDF captures formatting information from a variety of desktop publishing applications, making it possible to send formatted documents and have them appear on the recipient's monitor or printer as they were intended.

XML Extensible Markup LanguageThe Extensible Markup Language (XML) is the universal format for structured documents and data on the Web.

4S Swift Search Software and Services