Integration of a Multilingual Keyword Extractor in a Document Management System

Andrea Agili[*], Marco Fabbri[*], Alessandro Panunzi[+], Manuel Zini[*]
[*]DrWolf s.r.l., [+] Dipartimento di Italianistica - Università di Firenze
[*]Via Ponte alle Mosse 43/a Firenze, [+]piazza Savonarola 1 Firenze
E-mail: andrea.agili@drwolf.it, marco.fabbri@drwolf.it, alessandro.panunzi@unifi.it, mlzini@drwolf.it

Abstract

In this paper we present a new Document Management System called DrStorage. This DMS is multi-platform, JCR-170 compliant, supports WebDav, versioning, user authentication and authorization and the most widespread file formats (Adobe PDF, Microsoft Office, HTML,...). It is also easy to customize in order to enhance its search capabilities and to support automatic metadata assignment. DrStorage has been integrated with an automatic language guesser and with an automatic keyword extractor: these metadata can be assigned automatically to documents, because the DrStorage's server part has benn modified to allow that metadata assignment takes place as documents are put in the repository. Metadata can greatly improve the search capabilites and the results quality of a search engine. DrStorage's client has been customized with two search results view: the first, called timeline view, shows temporal trends of queries as an histogram, the second, keyword cloud, shows which words are correlated and how much are correlated with the results of a particular day.

1. Document Search Problem

It is common experience to store lots of heterogeneous documents in various folders on a local hard disk or in a network file system. Then, after some time, it is also common to have the need to retrieve some of those documents, but not to remember which their location was.
It is also a common need to assign metadata to stored documents. Metadata generally provide useful information and are used to obtain better search results. The drawback is that the manual metadata assignment is a boring and error prone task, so automatic metadata assignment would be a great benefit.

The market has supplied several and various tools which help in documents searching. Some of them are very rough (e.g. MS Windows search into file), other are more sophisticated (Google Desktop, MS Windows Desktop Search, Beagle, Spotlight, Strigi...).

Anyway, they lack some features that may improve query results relevance.

First of all, they generally treat documents (files, email, web histories or whatever) as they are, so it's not possible to assign metadata and, consequentially, it's neither possible to search on them, nor to use them for results ranking.

Moreover, they are designed to work on a single machine, they don't support documents sharing and it is not easy to search on them from various hosts in a net (Beagle supports NFS, but indexes are stored on a single machine and searches are possible from that machine only).

2. Document Management Systems

To address this kind of problems a new kind of software has been developed: Document Management Systems (DMS). A typical DMS is capable to:

- manage a network repository allowing users to store documents via a number of way (Server Message Block SMB, WebDAV, dedicated clients, ...);
- automatically extract low level metadata (size, timestamps,...) and the text from a ever growing number of document formats (MS Office, Adobe PDF, HTML, XML,...);
- store and index documents, metadata and texts;
- make queries on documents contents and metadata;
- manage user, authentication, roles, authorizations.

Some of them are also capable to do versioning and to create and apply workflow on documents.

In this paper we are going to present a new DMS, called DrStorage, which features all the above mentioned capabilities, but also allows integration with linguistic technologies in order to automatically extract and assign metadata from inserted documents.

3. DrStorage

DrStorage is a DMS written entirely in Java, so it is natively multiplatform (runs smoothly on MS Windows, GNU/Linux or Mac OS X systems). It is based on the open source project JLibrary, but it has been reengineered in order to be easily extensible in searching capabilities, but also to create clients applications with customizable search feature and result views. JLibrary, in turn, exploits Apache Jackrabbit which is a fully conforming implementation of the Content Repository for Java Technology API (JCR).

Among the main features of DrStorage are:

- compliance with JSR-170, the Java Standard Request for Content Repository;
- support for WebDAV;
- support for editing lots of documents formats such as MS Office formats, Adobe PDF, HTML, plain text;
- documents versioning;
- support for high-level metadata such as language, keywords, categories.

4. Search techniques

Searching on full text indexes is a mature and largely used feature and its implementation can be found in hundreds of desktop applications and search engines. Many kinds of queries have been developed and are present as features in several applications: boolean queries, range queries (e.g. using date and time values), fuzzy queries (e.g. Google’s ‘Did you mean…’), proximity queries, regular expression queries. It can be stated that the problem of retrieving a set of documents, given a query, have been solved many years ago and the solution is largely adopted. What it's already far from being achieved is to get what it's more important at the top of the results list. In other words, what's now really interesting is to rank results, not merely to get them. Many techniques have been developed to obtain better rankings: for the web documents that are linked on the net, Page Rank is one of the most famous. For texts, keywords can be considered a valid support: the basic idea is that if a word in a query is also a document keyword, then that document should receive a high rank, because the keyword is a highly descriptive word of the document content.

Obviously, for a large amount of documents, this approach is feasible only if the keyword extraction process can be done automatically.

5. Implementation features

DrStorage implements a multilingual keyword extractor developed by LABLITA (Italian Department, University of Florence) (Panunzi, Fabbri, Moneglia, 2005). The algorithm behind the implementation exploits well-known statistical techniques like TF.IDF, together with automatic PoS-Tagging and linguistic filtering rules. The algorithm has been proved to provide good results for the following European languages: English, French, German, Italian and Spanish.

The algorithm is also capable of extracting multi-terms keywords which have been proved to be much more descriptive of the document content than mono-term ones (Panunzi et al., 2006).

LABLITA's keyword extractor algorithm has been developed inside European Project AXMEDIS (Automating Production of Cross Media Content for Multi-channel Distribution, IST-2-511299) and an implementation in C++ is now part of the project content processing infrastructure.

The implementation exploits a set of linguistic statistical references and an external part of speech tagger and lemmatizer: a complete description of the implementation can be found in (Panunzi et al., 2006) and (Panunzi, Fabbri, Moneglia, 2006).

DrStorage leverages a new Java implementation of the algorithm with updated statistics and linguistic rules for the supported languages.

DrStorage implements also a language guesser, based on the open source library NGramJ for automatic extraction of documents language.

Every document that is stored in DrStorage's repository passes through a customizable processing pipeline that can assign metadata, reformat the document, extract further information. The language guesser and the keyword extractor can be part of the process, so text documents that are processed by the pipeline have their language extracted and a list of multi-term keyword assigned. All this process goes with no user interaction, but the starting command of putting the document in the repository.

This way, is very simple to create a repository with high-level metadata, such as language and keywords: in figure 1, we can see a very simple repository built by importing a mailbox; the repository is presented with a common filesystem tree layout which can be explored in the usual way, by opening folders and subfolders and by double clicking on documents to access them.

With DrStorage it's now easy to search on documents. In figure 2 it's shown a typical search dialog with which various kind of queries can be performed; it's possible to filter documents of a selected language only, to exploit keywords just by searching on them (resulting in a very fast query with almost no "noise" results) or by using them to boost the rank position of certain items in a result set. Figure 3 is given as an example to show how results are presented by the interface: the example show results obtained querying the simple word 'mobile' (as shown in figure 2) on the mailbox repository.

6. Customization

DrStorage is also easy customizable to display search results in other ways than just listing. These new results views can help the user to infer new information about the documents in the repository.

For example, DrStorage provides a timeline view (figure 4) that exploits both search results and documents time metadata, in order to see temporal trends of queries or for keywords. This view is displayed as a bar chart with dates on the y-axis and search results number on the x-axis; the view can be useful to infer information on documents for which the time metadatum is important, such as news or Public Administration documents or newsgroups comments; at a glance, it is possible to understand when newspapers had written about a certain event (verifying how many repercussions the event has had in the next days), or if a certain topic has been discussed in web forums (maybe for marketing monitoring), or to see if a Public Administration had treated an argument and with which "importance" (more documents about an argument, more importance the PA gave to it).

From this view it's simple to obtain documents of a certain day, by simple double-clicking on the relative bar.

Another useful result view is the keyword cloud view: given a result set, the view shows (in a format very similar to common tag-cloud) the more frequent keywords of the documents in the result set. Keyword in this case are obtained with a simple TF.IDF algorithm which compares only the words contained in the documents relatives to the day against the words contained in the whole repository; the cloud may demonstrate that documents that have been found with a particular query, convey also information regarding other topics, so showing correlations that might not be evident without a deep analysis of each document in the result set. The keyword cloud can be displayed by simply right-clicking on a day-bar on the timeline view and choosing 'Keyword Cloud' from the pop-up menu. In figure 5, we can see which are the most important words in the document of July 9th that contain the word 'mobile': at a glance it is possible to realize what those documents deal with.

7. References

Apache Jackrabbit http://jackrabbit.apache.org/

Beagle http://beagle-project.org/Main_Page

DMS http://en.wikipedia.org/wiki/Document_management_system

DrStorage http://www.drwolf.it

Google Desktop http://desktop.google.com/

JLibrary http://jlibrary.sourceforge.net/

JSR-170 http://www.jcp.org/en/jsr/detail?id=170

NFS http://tools.ietf.org/html/rfc3530

NGramJ http://ngramj.sourceforge.net/

PageRank USPTO Patent http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=7058628.PN.&OS=PN/7058628&RS=PN/7058628

Panunzi, A. Fabbri, M. Moneglia, M. Zini, M. (2006) Multi-Term Keywords for Indexing Multilingual Textual Repositories: Developing Language Resources and Algorithms, in *Proceedings of AXMEDIS 2006 Conference*, December 13-15, Leeds, UK.

Panunzi, A. Fabbri, M. Moneglia, M. (2006) Integrating Methods and LRs for Automatic Keyword Extraction from Open-Domain Texts, in *Proceedings of LREC 2006 Conference*, May 24-26, Genova.

Panunzi, A. Fabbri, M. Moneglia, M. (2005) Keyword Extraction in Open-Domain Multilingual Textual Resources, in *Proceedings of 1st International Conference on Automated Production of Cross Media Content for Multi-channel Distribution* (AXMEDIS 2005), IEEE Computer Society.

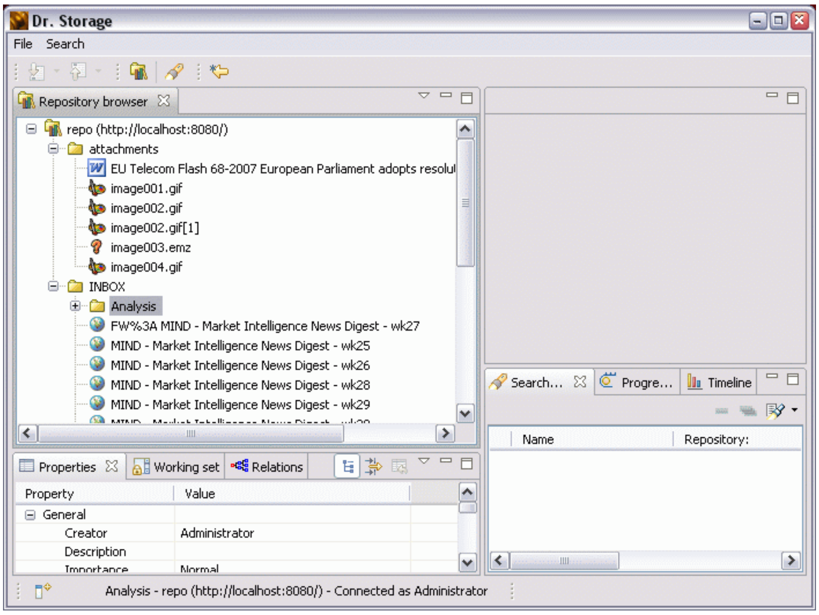
SMB http://en.wikipedia.org/wiki/Server_message_block

Spotlight http://docs.info.apple.com/article.html?artnum=304778-en

Strigi http://strigi.sourceforge.net/

WebDAV http://www.webdav.org/specs/

Windows Desktop Search http://www.microsoft.com/windows/products/winfamily/desktopsearch/default.mspx

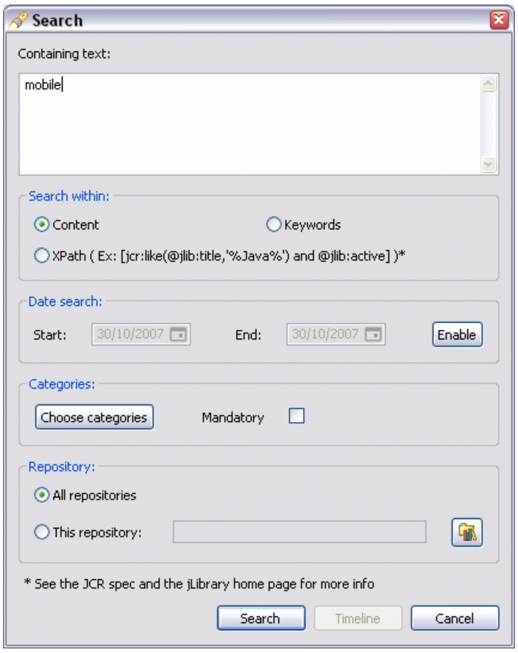


repo (http://localhost:8080/)
attachments
EU Telecom Flash 68-2007 European Parliament adopts resolul
image001.gif
image002.gif
image002.gif[1]
image003.emz
image004.gif
INBOX
Analysis
FW%3A MIND - Market Intelligence News Digest - wk27
MIND - Market Intelligence News Digest - wk25
MIND - Market Intelligence News Digest - wk26
MIND - Market Intelligence News Digest - wk28
MIND - Market Intelligence News Digest - wk29

Property	Value
General	
Creator	Administrator
Description	
Importance	Normal

Analysis - repo (http://localhost:8080/) - Connected as Administrator

Figure 1: simple repository view.



Containing text: mobile

● Content ○ Keywords
○ XPath (Ex: [jcr:like(@jlib:title,'%Java%') and @jlib:active])*

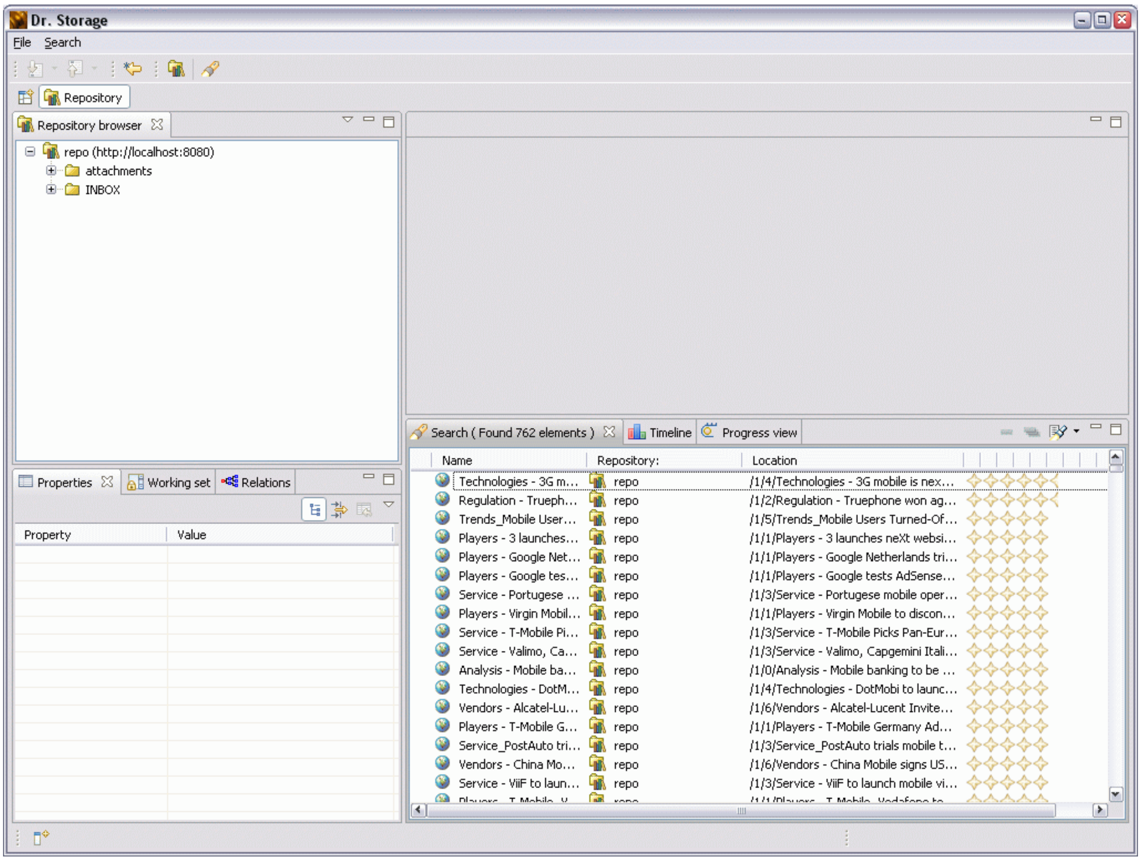
Start: 30/10/2007 End: 30/10/2007

Mandatory □

● All repositories
○ This repository:

* See the JCR spec and the jLibrary home page for more info

Figure 2: search dialog



Name	Repository:	Location
Technologies - 3G m…	repo	/1/4/Technologies - 3G mobile is nex…
Regulation - Trueph…	repo	/1/2/Regulation - Truephone won ag…
Trends_Mobile User…	repo	/1/5/Trends_Mobile Users Turned-Of…
Players - 3 launches…	repo	/1/1/Players - 3 launches neXt websi…
Players - Google Net…	repo	/1/1/Players - Google Netherlands tri…
Players - Google tes…	repo	/1/1/Players - Google tests AdSense…
Service - Portugese …	repo	/1/3/Service - Portugese mobile oper…
Players - Virgin Mobil…	repo	/1/1/Players - Virgin Mobile to discon…
Service - T-Mobile Pi…	repo	/1/3/Service - T-Mobile Picks Pan-Eur…
Service - Valimo, Ca…	repo	/1/3/Service - Valimo, Capgemini Itali…
Analysis - Mobile ba…	repo	/1/0/Analysis - Mobile banking to be …
Technologies - DotM…	repo	/1/4/Technologies - DotMobi to launc…
Vendors - Alcatel-Lu…	repo	/1/6/Vendors - Alcatel-Lucent Invite…
Players - T-Mobile G…	repo	/1/1/Players - T-Mobile Germany Ad…
Service_PostAuto tri…	repo	/1/3/Service_PostAuto trials mobile t…
Vendors - China Mo…	repo	/1/6/Vendors - China Mobile signs US…
Service - ViiF to laun…	repo	/1/3/Service - ViiF to launch mobile vi…

Figure 3: search result example

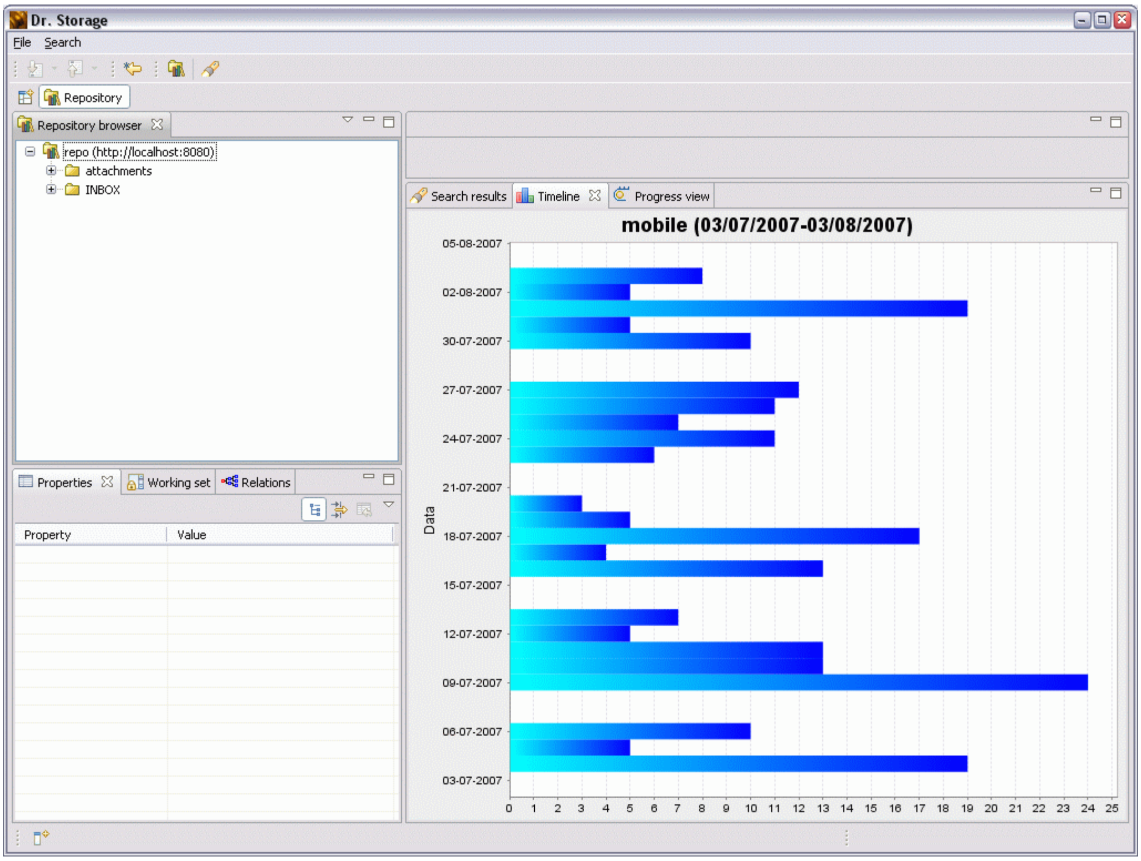


Figure 4: timeline view of the results on the query 'mobile'

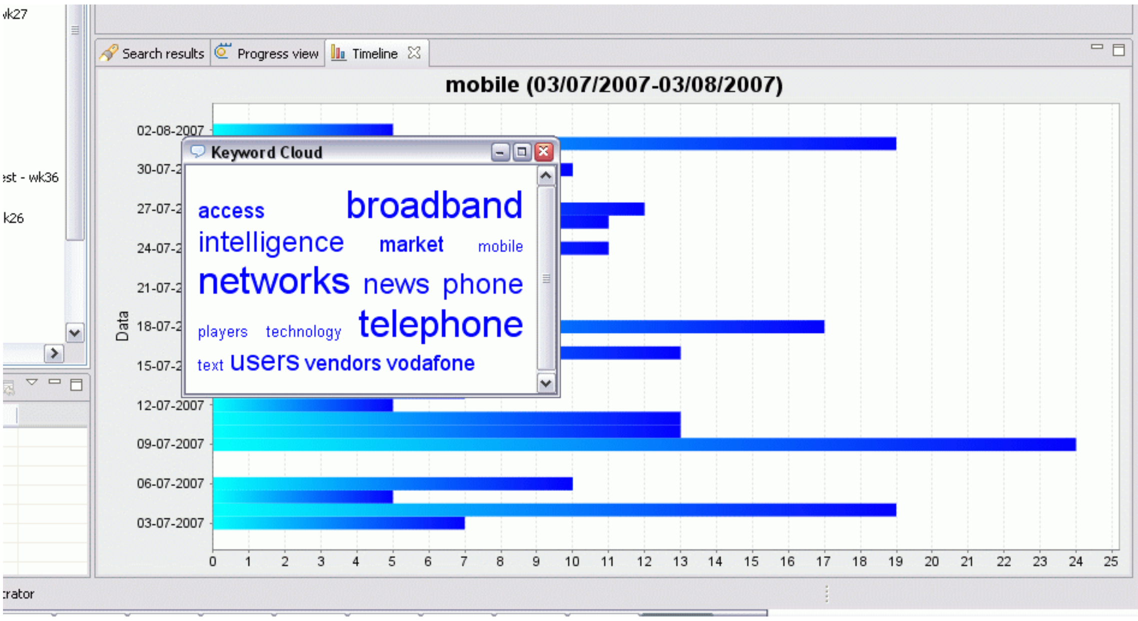


Figure 5: keyword cloud example