Alfarrábio: Adding value to an heterogeneous site collection

José João Dias de Almeida
Universidade do Minho
jj@di.uminho.pt

Pedro Rangel Henriques
Universidade do Minho
prh@di.uminho.pt

Jorge Gustavo Rocha
Universidade do Minho
jgr@di.uminho.pt

Alberto Simões
Universidade do Minho
albie@alfarrabio.di.uminho.pt

Abstract

The Web is probably the largest collection of digital resources ever built, and it maintains it's continuous growth. But this large collection is not a digital library, according to the most accepted sense, since it lacks some fundamental mechanisms.

While it is possible to create and maintain a structured site, with powerful cataloging mechanisms and information retrieval, in a way that can be considered as a digital library, the task of grouping several heterogeneous sites with the same purpose is a more complicated problem, and several approaches can be taken.

In this article, we report our effort to develop information extraction, cataloging and browsing mechanisms to a set of heterogeneous cultural sites, to be regarded, from the user's point of view, as a single digital library. Our case study is based on the Alfarrábio project, a cultural cooperative to maintain and support the publication of cultural resources on the Web, both by individuals and non-profit organizations.

The cataloguing task of every relevant document (HTML pages are just specific instances of documents) was one of the main concerns of the project; it is supported by automatic tools. We build a semantic net of all related terms (according to predefined and user's defined relations) crossing all Alfarrábio resources: whole sites, HTML pages, images, travel maps, sound files, XML documents, relational data, etc.

This is quite different from the approaches taken by the so-called portals, where information within different channels is not crossed, and resources (and whole sites) are only cataloged in broader terms.

Due to the lack of human resources, we developed a set of automatic cataloging tools and dynamically generated browsing tools. The participation of each site creator in the cataloguing of his own information is supported and encouraged.

In this article we present and explain the developed tools for information extraction, for cataloging maintenance, and for conceptual browsing.

Introduction

The Internet is well recognized for the dramatic increase of information availability, but this flood of information does not mean a proportional knowledge increase. Many Web publishers, and the large majority of users are struggled by the chaotic nature of the Web. This problem is only partially solved by general portals like Yahoo, and several initiatives are being taking all over the world to address it .

In this paper, we describe our approach to build a new layer over web resources, providing a suitable information structure and enhanced browsing capabilities.

Our case study is Alfarrábio (http://alfarrabio.um.geira.pt), under which several cultural sites are accessible. Apart from the importance and value of each one, we would like to make available to users an additional value, greater than just the union of all the Alfarrábio sites.

We will describe the Alfarrábio project presenting the main design decisions, the classification structure used, the catalogue format and construction, and we also describe the basic tools referred.

About Alfarrábio

Alfarrábio project was born from the need to store in a common area several individual Internet initiatives developed by various persons (the alfarrabists) and somehow related to culture.

The project started as a bottom-up process, to prove that it is possible to build a useful archive from heterogeneous non-standardized volunteer initiatives.

At moment, Alfarrábio contains sites about:

- music (including lyrics, scores, karaoke)
- cultural events of Braga
- Portuguese-Chinese dictionary
- Portuguese literature archives
- monographes and geographic information
- children stories
- scientific clubs
- red-cross organizations
- oral tradition and life stories
- sport
- short Portuguese lessons
- ...

Alfarrábio is served by a Linux system, using free software.

Design goals

Experience shows that it is not easy to impose standards. Alfarrabists have very different background, use different tools to produce their sites, live in different places... Typically, they have their professional activities and develop Internet information about culture as a hobby.

In order to make a whole from the set of individual Internet sites, it was decided to build an extra level:

- to give a common view (a rich catalogue using a common format)
- to make as many as possible (conceptual) connections between documents

For each site, we must have a catalogue entry describing the site in general, and a catalogue describing each (useful) sub-document.

To establish as many relations as possible, we decided to build and use:

- a rich classification structure (thesaurus, a semantic network)
- a large set of predefined relations, to classify each document (relating the documents to the classification structure)

and to develop a set of tools to make the classification as automatic and powerful as possible:

- tools to help in the process of building the catalogue
- tools to build conceptual navigation from thesaurus and catalogue
- tools to search, taking profit of the conceptual structure

One of the main goals was to guarantee that the information produced, would be available for a long period.

The more we got envolved in the project the more we like it... and the project became a preferential case study to test experimental ideas.

The tools developed were the result of applying a programming approach to a librarian problem. With the growth of the information amount, we proved the usefulness and usability of that set of tools.

Developed tools

After discussing some archive decisions[**?**], we present the tools built. Those tools are in different stages of development; some of them are generally reusable, but others are not yet robust enough to be useful in other projects.

The tools can be grouped in two classes, one related to the thesaurus, and the other addressing the catalogue.

Tools to build and use the thesaurus:

- `alfaThe.pm` – Perl Module for thesaurus processing, translation, completion (based on a set of rules) and use
- `athesaurus` – CGI script for conceptual navigation and searching. Parameters: a thesaurus and a catalogue. Uses: `alfaThe.pm`, `XML::DT`, `Mapit`.

Tools to aid catalogue construction:

- `getaa` - get catalogue information from a site index
- `getmp` - get catalogue information from metadata contained in a HTML page

Paper structure

In section , the catalogue contents, format, and construction is discussed.

In section , the thesaurus definition, properties and processing tools is described.

In the usual section of conclusions, the main results are presented.

The Catalogue

In order to have a common format, it was decided to use a XML format following a local catalogue DTD. In a simple way a catalogue is a set of document entries.

The catalogue can have document entries (`doc` tag) of different complexity and richness. Each entry contains title, author and their roles, urls, description, dates, small images, ..., and relations to terms of the thesaurus.

In order to make the process of rebuilding the catalogue, the catalogue is distributed by different files. This way, tecnologically advanced alfarrabists can be the owner of a catalogue file for (parts of) his site. Remember that we catalogue documents, and not just sites. Some sites have thousands of document entries.

To illustrate the idea, consider the following two examples of catalogue components, corresponding to one site include in alfarrábio. The first one is about the site Cancioneiro (Cancioneiro is a archive of music with near 80 musics in various shapes, mantained by Domingos Morais). The second is about a document contained in Cancioneiro.

```
<doc>
  <title xml:lang="pt">Cancioneiro</title>
  <resource xml:lang="pt">
         http://alfarrabio.um.geira.pt/cancioneiro
  </resource>
  <description><pre>
    .  Introduction
    .  Section 1: Children Songs in Portuguese
    .  Section 2: Songs in Portuguese</pre>
   <p> This set of scores is a ...  </p>
  </description>
  <author email="dmorais@ip.pt"
         role="Selection, karaoke and notes">
     Domingos Morais         </author>
  <relations>
    <rel type="IOF">book</rel>
    <rel type="IOF">arquive</rel>
    <rel type="POF">alfarrábio</rel>
    <rel type="ABOUT">music</rel>
    <rel type="ABOUT">Portuguese</rel>
    <rel type="ABOUT">karaoke</rel>
    <rel type="ABOUT">score</rel>
    <rel type="ABOUT">poem</rel>
  </relations>
</doc>

<doc>
  <title xml:lang="pt">Lá vem a nau Catrineta</title>
  <resource>
    http://alfarrabio.um.geira.pt/cancioneiro/110.html
  </resource>
  <description>score, lyrics and karaoke of
      "Lá vem a nau Catrineta"
  </description>
  <author email="dmorais@ip.pt"
         role="karaoke, partitura">
       Domingos Morais </author>
  <relations>
    <rel type="IOF">score</rel>
    <rel type="IOF">music</rel>
    <rel type="POF">Cancioneiro</rel>
    <rel type="IOF">karaoke</rel>
    <rel type="IOF">rimance</rel>
    <rel type="ABOUT">devil</rel>
    <rel type="ABOUT">boat</rel>
  </relations>
</doc>
```

In the following examples, we also have included XML tags to refer images (`icon`) and to geo-reference the document in a map (`where`).

```
<doc id="136">
 <resource>/alfabraga/photos/P0000146.JPG</resource>
 <icon>/alfabraga/photos/thumbs/P0000146.JPG</icon>
 <title>Arcada</title>
 <author>Augustsson</author>
 <relations>
  <rel type="geo">Braga</rel>
  <rel type="iof">Praça</rel>
  <rel type="iof">foto</rel>
  ...
 </relations>
 <where mapa="braga1" x="0.4724324" y="0.6107611"/>
</doc>
```

It is possible to include other fields in the catalogue entries. That is usefull for documents with very specific data. Some of them will be included in the catalogue DTD, in the near future.

Extracting information from sites and documents

There are several types of sites and several ways to build each one.

In a simple way:

- a site can have strong internal structure – In this case it can be possible to build an export funtions that writes a catalogue file.
- a site is a archive – in this case the author, often builds indexes: a view over the set of subdocument in the archive. The catalogue, or parts of it, can be obtain by adding generic information of th site with specific information of each index entry. See section

- a site has a clear set of documents – in this case we can try to extract meta-information from the document meta-information (if available).
- a site can have no clear structure – in this case the catalogue can be written with the help of interactive tools

`getaa` – **extract catalogue from site indexes.**

`getaa` is meant to help in the task of extracting (poor) catalogues from a index of a cooperating site.

`getaa` is a perl script that receives a catalogue entry skeleton, a url of an index of a site, and optionally a set of configuration parameter and returns a catalogue.

`getmd` – **extract catalogue from metadata in documents.**

HTML has mechanisms to store catalogue information in the header of the HTML document.

When a document/subdocument has a rich metadata definition (Ex. following Dublin core [**?**]), `getmd` extracts that metainformation and helps in the task of editing it.

Unfortunately, in certain environments this tool can not be used. We noticed that some interactive tools delete the non visible information of the documents!

Generating catalogue from rich archive sites

The catalogue's DTD is available. So each alfarrabist can produce, or edit, his own catalogue, either manually or automaticaly.

Some of the sites have internal databases, or have a rich structured definition in XML or similar. Tipically, the sites have translation layers to produce HTML files. In this case it is easy to generate their catalogue file by writing an extra translation layer to generate the XML catalogue, according to the given DTD.

The thesaurus

The thesaurus (librarian view) used in Alfarrábio complies with thesaurus ISO.

In a simple way, the thesaurus defines a closed set of active terms to be used in classification activities, and a set of relations between them.

$$
\begin{array}{lcrcl}
thes & = & terms & : & semNet \times \\
 & & prop & : & Tprop \\
semNet & = & terms & : & set(term) \times \\
 & & edges & : & set(edge) \\
edge & = & ori & : & term^{\star} \times \\
 & & r & : & Rel^{\star} \times \\
 & & dest & : & term
\end{array}
$$

Each relation (Rel) has mathematical properties that allow to make a certain amount of inference.

The thesaurus rich relation set

The basic relations used in the alfarrábio's thesaurus are:

- BT/NT – (Boarder term/narrower term) (standard in librarian studies) (antisimetric, antireflexiv, transitive)
- USE/USES – (use instead/used for) preferential term(standard in librarian studies)
- POF/HAS – (Part of/has) relation (anti-simetric, anti-reflexive, transitive)
- IOF/INST – (Instance of/instances) (anti-simetric, anti-reflexive, transitive)
- makes/by – Relation between the author and his work (anti-simetric, anti-reflexive)
- SN – Scope note
- RT – Related term (simetric)

The thesaurus contains term about: subject, geographic elements, temporal epochs, type of documents, people, etc.

In the present version, the thesaurus of alfarrábio the contains near 500 active terms.

The thesaurus processing tools

The `AlfaThe.pm` perl module

The module `AlfaThe.pm` contains many functions to process the thesaurus. This module is used in every tool that processes thesaurus.

The basic functionality of the module is:

- functions to import/export thesaurus
  $import\_txt : ISOtheTxt \longrightarrow thesaurus$
  $import\_txt(file) \stackrel{\text{def}}{=}$
  loads the thesaurus from a ISOthesaurus file
  $dump : thes \longrightarrow ISOtheTxt$
  $dump(t) \stackrel{\text{def}}{=}$
  writes a ISOthesaurus file from the thesaurus
  $toXml : thes \times table \longrightarrow XMLtxt$
  $toXml(t, details) \stackrel{\text{def}}{=}$
  writes a XML file from the thesaurus
  $toTex : thes \times table \longrightarrow LaTeX$
  $toTex(t, details) \stackrel{\text{def}}{=}$
  writes a LaTeX file from the thesaurus

- functions to ask information about a term (or set of terms)
  $terminfo : thes \times Term \longrightarrow Rel \hookrightarrow Term^{\star}$
  $terminfo(t, ter) \stackrel{\text{def}}{=}$
  All the information about the term ter
  $transClosure : thes \times set(Rel) \times term \longrightarrow set(term)$
  $transClosure(t, rs, ter) \stackrel{\text{def}}{=}$
  the terms acessible from term using edges rs
- functions to define mathematic properties of the relations and functions traversal the all struture
  $Tprop = invers : Rel \hookrightarrow Rel$
  makes the thesaurus completion

  $depth\_first : thes \times term \times ANY \times set(Rel) \longrightarrow$
  $depth\_first(T, t, visit, R) \stackrel{\text{def}}{=}$
  depth first visit of T using relations in R

The `athesaurus` cgi conceptual browser

With `athesaurus` cgi, we can:

- browse the thesaurus conceptual structure and define a current term
- search in the catalogue for documents that contain a pattern and that are related with a conceptual term (by default: the current term).

Search function has two parameters: a pattern and a thesaurus term. A document is selected if it matches the search pattern and belongs to the transitive-closure of the term. The theaurus term used in search, can be provided in the search expression or default to the current term.

In figure 1 we see the result of searching for documents which are houses and match "brasileira" (brasilien). The expression used was `brasileira:house`. Note that many documents match the pattern "brasileira" (ex: brasilien music).

The search engine evaluates the transitive closure of "house" which contains the term "coffee-house". The returned documents, besides containing "brasileira" are classified as coffee-houses.

The result presents two kinds of information: a view of the thesaurus centered in the term "casa" (=house) and a view of the catalogue filtered by the search expression.

`athesaurus` builds a log file that can be use in the task of improving the thesaurus: when a concept is not found it can be addes as a non-preferencial term (a synonimous of a thesaurus term).

Using maps for catalogue navigation

Another way to navigate over catalogue information is using a map. Consider a city, and a catalogue of the monuments: on one hand the map can be used to show where a particular monument is; on the other hand, you can navigate over the map and click on it. If there is any document near the clicked point, a list of documents will be returned.

This can be done easily as was shown in the third example in section . First, it is necessary to add information, in each document, about the coordinates in a map, where it fits. This can be done using a XML tag like this:

```
<where map="braga1" x="0.47243" y="0.61076"/>
```

The `map` attribute has the map name, and the `x` and `y` attributes contain pixel coordinates in the map image file. The visualization of the point in the map, is done by a script that, using a `gif` map file, draws a dot in the selected coordinates and shows it in the browser.

To go in the oposite direction, we use a script that receives the coortinates of point on a gif, and search in the catalogue for documents with coordinates in the neigbourhood ($x \pm \epsilon$ and $y \pm \epsilon$ where ϵ is a value to consider incorrections).

In the Alfarrábio project we developed some perl script to produce map browsers: `mapit`. It receives a big gif and cuts it on small pieces (small gifs that can be viewed in a normal browser window). It creates a script, too, to navigate over the map (pan from small gif to small gif, and see a small version of the big map), and to click and send the coordinates to another script, that process it and shows a listing os the catalogue entries that match with the coordinates.

Conclusion

In this paper, we have discussed our approach to enhance a set of cultural sites with an additional layer to provide users with a more powerful digital archive, adding to the whole archive some extra knowledge.

$(knowledge(digarchive) > \sum knowledge(theparts))$.

Some major achievments of the Alfarrábio project:

- The selection of the relevant material for cataloguing and retrieving is based on "what is most important" rather than "everything that is digitally available",

- The creators are not forced to use any specific standard,
- A space for creators participation and support is provided as it is a must,
- However all the tools tends to be as automatic as possible (require few user interaction),
- Only open standards and free software tools were used.

All parts of the system are scalable and thus usable for smaller initiatives then Alfarrábio, and for larger ones; differences will reside in the size of the catalogue (a function of the number of the documents and level of subdocuments considered) and its grain (the amount information in each entry).

References

[1] Serge Abiteboul. Querying semi-structured data. *Proc. of Int. Conf. on Database Theory (ICDT), Delphi, Greece*, 1997.

[2] M. Balabanovic and Y. Shoham. Content-based, collaborative recommendation. *Communications of the ACM 4 n. 3 66-72*, 1997.

[3] Roy Fielding et al. Web-based development of complex information products. *Communications of the ACM, 41-8*, Aug 1998.

[4] D. Florescu, A. Levy, and A. Mendelzon. Database techniques forthe www: A survey. *SIGMOD Record 27-3*, Sep 1998.

[5] V. Gudivada, V. Raghavan, W. Grosky, and R. Kasanagottu. Information retrieval on the www. *IEEE Intelligent Computing*, 1997.

[6] Naveen Ashish Craig Knoblock. Wrapper generation for semi-structured internet sources. *SIGMOD Record, 26-4*, Dec 1997.

[7] S. Nestorov, S. Abiteboul, and R. Motwani. Inferring structure in semistructured data. *SIGMOD Record 26-4*, Dec 1997.

[8] D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. Querying semistructured heterogeneous information. *International Conf. on Deductive and Object-Oriented Databases*, 1995.

Screenshots

Figure 1: athesaurus: search for "brasileira:casa"