Sistemas de código abierto para la creación, mantenimiento y aprovechamiento de bibliotecas digitales (SCABD)
TIN2006-15071-C03

Rafael C. Carrasco Jiménez*
Dep. de Llenguatges i Sistemes Informàtics, Universitat d’Alacant

Pablo L. de la Fuente Redondo†
Dep. de Informática, Universidad de Valladolid

Nieves R. Brisaboa‡
Dep. de Computación, Universidade da Coruña

Abstract

This project aims to generate new techniques and open-source software tools that help in the production and maintenance of large digital libraries and allow for more intensive exploitation of their contents. In particular, software has been implemented to enable the linguistically-enriched exploitation of digital libraries and, to some extent, to improve the transcription quality of digital content and to allow cheaper production of digital texts and metadata. Also, new methods to index and retrieve documents with structural markup (even in compressed form) in large collections have been explored. Additionally, new techniques to the creation and use of educational resources have been tested.

Keywords: digital libraries, open-source code.

1 Project objectives

The main objectives of this project are:

1. Improving correction algorithms for OCR from a large collection of digitized texts supervised by experts.
2. Creating tools for the structural markup of digitized texts.
3. Designing efficient methods to compress and later search compressed texts, especially for highly inflective languages.

*E-mail: carrasco@dsi.ua.es
†E-mail: pfuente@infor.uva.es
‡E-mail: brisaboa@udc.es
4. Creating new methods to improve the indexing and retrieval of structured information.

5. Implementing linguistically intelligent searches in texts written in languages that contain many forms that are not found in dictionaries (inflected verbs, feminines and plurals, etc.);

6. Creating tools to allow new educational uses of digital libraries.

These objectives generate the following tasks:

- **Module 1: Improvement of the algorithms for the correction of the results of automatic transcription.**
  - Task 1.1: Implementation of unsupervised methods.
  - Task 1.2: Implementation of supervised methods; comparison with unsupervised.
  - Task 1.3: Integration of methods based on probabilistic finite-state machines.

- **Module 2: Tools for the structural markup of digitalized texts**
  - Task 2.1. Automatic preliminary markup based on automatic transcription.
  - Task 2.2. Adaptive tools for assisted markup.

- **Module 3: Indexation of texts with compression**
  - Task 3.1. Gathering of a corpus of texts in Romance languages.
  - Task 3.2. Test compression techniques with direct search in Romance corpora.
  - Task 3.3. Adaptation of the text compression techniques to Romance languages.
  - Task 3.4. Implementation of indexes and a suffix arrays for compressed text.
  - Task 3.5. Empirical validation of the inverted index and the suffix array.

- **Module 4: Indexation and retrieval of structured information.**
  - Task 4.1. Design and implementation of a model for the retrieval of information with structure.
  - Task 4.2. Analysis, design and implementation of a compression and indexation system on texts considering their structure.
  - Task 4.3. Study of the representation of the structure in query interfaces.
  - Task 4.4. Evaluation of the systems developed in tasks 4.1, 4.2 and 4.3.

- **Module 5: Linguistic processing.**
  - Task 5.1. Automatic generation of search suggestions.
  - Task 5.2. Searches with morphological analysis and categorial disambiguation.
  - Task 5.3. Assisted searches with shallow parsing and multi-word expressions.

- **Module 6: Tools for educational purposes.**
  - Task 6.2. Implementation of the specification in different components.
- Task 6.3. Definition of an ontology of learning objects.
- Task 6.4. Implementation of a content management system.

- Module 7: Publication of the developed tools as open source.
  - Task 7.1. Implementation of the open-source packages
  - Task 7.2. Generation of documentation. Description of tasks

The distribution of tasks among the partners and periods is shown below.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Partners</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
<th>T4</th>
<th>T5</th>
<th>T6</th>
<th>T7</th>
<th>T8</th>
<th>T9</th>
<th>T10</th>
<th>T11</th>
<th>T12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1 Unsupervised</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2 Supervised</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.3 Finite-state</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Markup</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1 Incremental</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1 Corpus</td>
<td>UdC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.2 Testing</td>
<td>UdC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.3 Modification</td>
<td>UdC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.4 Implement</td>
<td>UdC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5 Validation</td>
<td>UdC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1 Models</td>
<td>UVa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.2 Compression</td>
<td>UVa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.3 Interface</td>
<td>UVa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.4 Evaluation</td>
<td>UVa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1 Suggestions</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2 Morphology</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.3 Syntax</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1 Specification</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.2 Implement.</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.3 Ontology</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.4 CMS</td>
<td>UA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1 OpenSource</td>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2 Achievements

2.1 Advances in the tasks

- Module 1: A software has been implemented to align OCR text (in RTF format) with supervised text (in XML format). Special problems, such as the movement of footnotes from their physical to their logical position, received careful treatment.

- Module 2: Our search for methods to automate structural markup produced results in two unforeseen but highly relevant aspects: the efficient storage of structural information as finite-state machines (DTA), and its usage to create schemes for (minimal perfect hashing). These results were published[24] in a high impact journal, Algorithmica, and at two international conferences: CIAA 2007[23], and IIS 2008[27]. Additionally, a new journal paper[22] is under revision at Theoretical Informatics and Applications

- Module 3: All the objectives in this module were reached and also advances in new methods for text indexing and compressing were made. Additionally, we have started to study compression and indexing on different data types such as cartography[37] or graphs representing the Web structure or topological relationships among geographic objects in cartographic databases.
Task 3.1: we created huge digital libraries in Galician and Spanish that can be accessed on Internet and described them in journals listed in the JCR. Tasks 3.2 and 3.3: we obtained remarkable results in both static (3.2) and dynamic (3.3) compression and published them in high impact journals: Information Retrieval (IR) and Software: Practice and Experience.

Task 3.4: we implemented the word-oriented compressed suffix array published in the SPIRE conference, and also developed a new compressed self-indexed structure (called wavelet tree) which achieves better retrieval times than standard indexes and spaces about 35% to 40% of the original text, as published in SIGIR, the best international forum for information retrieval and submitted to ACM-TOIS and IR.

Module 4 (task 4.1): Our search for a model of information retrieval taking into account document structure has produced an architecture (Aqueducts) for XML Processing published in an international conference (ECSA2007). The use of Aqueducts in Services-Based Architectures (SOA) was presented in several conferences and published in the journal IEEE América Latina.

Task 4.2: We proposed new methods to compress XML documents: one (LZCS) which replaces repeated substructures with a backward reference and was published in the journal Journal of the American Society for Information Science and Technology (JASIST) and a second one in Information Processing and Management. We also implemented a XPath-like system to browse LZCS documents.

A new compression technique for natural language text collections that use a graph to model the text (edge-guided) has been published in two conferences. A methodology to automatically extract and resolve references to fragments of XML documents was published in Information Processing and Management.

Task 4.3: We have defined Electronic work (E-work) as a new electronic book metaphor based on logical structure that represents structure using a TEI Lite based scheme and uses Aqueducts to process the content and organize functionality and style of the services. The logical structure scheme is defined in a journal and different versions of the model and its application have been presented in conferences.

Additional results in module 4: We collaborated with Telefónica Investigación y Desarrollo (TID) in a project on contextual search applied to the Mobile Web environment. Results and a prototype have been presented or accepted at international conferences. A case study of the adaptation of interfaces has been published as a chapter in a book about User Interfaces.

Module 5 (task 5.2): Our research has produced a software module for the Lucene project, called apertium-morph, which provides a set of tools and Java classes that allow the Apache Lucene text search engine to use morphological information to index and search. To that end, the linguistic resources developed for the Apertium open-source machine translation platform are used to extract morphological information while indexing. This component was presented at the Workshop on Information Access to Cultural Heritage. The module can be downloaded from the sourceforge repository. The module has been downloaded more than 100 times since its publications in July 2008.
Task 5.3: The integration of linguistic tools to search and index texts in digital libraries has benefited from the development of new methods to train part-of-speech taggers. These results have been published as a PhD Thesis[70] and also as articles[72, 68] in an international journal, *Machine Translation* and in *Procesamiento del Lenguaje Natural*. The creation of modules to allow for syntactic search will benefit from the advances in the automatic detection of syntactic transfer rules published in the Theoretical and Methodological Issues in Machine Translation (TMI2007). conference[69].

- Module 6 (task 6.1): A new architecture for digital educational repositories has been proposed[35] and enhanced with new distributed search models[32]. These ideas have been extended to open content repositories[32]. Also tools for a richer exploitation of educational repositories have been implemented[57, 59].

Task 6.4: A tool has been implemented to help teachers to create video presentations using open-source tools with small requirements and high quality and the software has been presented at two international conferences[66, 67], IADAT 2007 and IATED 2007.

2.2 Project coordination

The project coordinated its goals by means of annual workshops and, as a result, some common publications arose: some between UA and UVa on bitexts compression[44, 2] and some between UVa and UdC about structure compression[4, 3, 5].

2.3 Difficulties and risk management

Scientific research is an intensive work and, therefore, research projects benefit most from those researchers which are full-time working on it. A harder situation is that from those who must share these duties with other tasks as university lectures and academic management. Progressively, universities are launching plans to evaluate internally the quality of the activities (research, teaching and management) run by their staff. This poses an additional source of stress on the staff, that must balance and split the amount of effort devoted to each task.

The more ambitious a project is, the higher are the risks that some objectives remain, partially or totally, unreached. However, as high-quality research must always have an speculative component, we consider that a partial achievement is not necessarily an indication of failure.

3 Performance indicators

3.1 Subproject 1: Herramientas lingüísticas y educativas

3.1.1 Trainees

Felipe Sánchez Martínez has completed his PhD Thesis[70] within this project. Two students, Francis Tyers and Xavier Ivars Ribes are working on the dissertation for Diploma de Estudios Avanzados (DEA) under our supervision.
3.1.2 Employees
Two people have been hired as técnico superior: Felipe Sánchez Martínez (June 2007–October 2008) and Xavier Ivars Ribes (November 2008–).

3.1.3 Publications

Articles in international journals

Articles in national journals

Lecture Notes in Computer Science

Communications in proceedings of international conferences


Communications in proceedings of national conferences

3.1.4 International cooperation

- Members of the group have participated in (and supervised) the MultiMatch research project (May 2006–October 2008), funded by the European Commission under the 7th Framework Programme as a Specific Targeted Research Project (contract No. 033104). MultiMatch developed a multilingual search engine specifically designed for access, organization and personalized presentation of cultural heritage information in digital libraries.

- Intense cooperation has been maintained with the Gdańsk University of Technology, leading to a number of common publications\cite{23, 27, 24, 22} and some publications where this project is acknowledged\cite{26}.

3.1.5 Technology transfer

- The software apertium-morph has been submitted to the IPR agency Registro Territorial de la propiedad intelectual de la Comunidad Valenciana. Solicitud: A-706-08, 28/10/2008 (Felipe Sánchez Martínez, Mikel Lorenzo Forcada Zubizarreta y Rafael Carlos Carrasco Jiménez).

- The cooperation with the Fundación Biblioteca Virtual Miguel de Cervantes has produced new developments based on the group’s know-how:

  1. The web service magister.cervantesvirtual.com which allows for the online creation of educational resources (in SCORM and XHTML standards) based on the contents of a digital library. Its implementation has been partially funded by the Spanish MCyT.

  2. A new architecture for a digital library that simplifies the creation of social (that is, web 2.0) services has been implemented and tested at the Fundación Biblioteca Virtual Miguel de Cervantes. This development was funded by the FBVMdC (01/04/2008–01/01/2009) and supervised by a member of the group (Pedro A. Pernías Peco).

- The OpenCourseWare project for the Spanish and Latin-American universities (OCW consortium), has been designed, implemented and supervised in its initial stage\cite{32} and, in the period 2006–2009 at the Universidad de Alicante, by a member in our group (Pedro A. Pernías Peco).

3.2 Subproject 2: Indexación y recuperación de información estructurada

3.2.1 Trainees

Miguel Ángel Martínez Prieto is working in his PhD thesis. We hope this it will be defended in 2010. Two students, José Manuel Cantera Fonseca and Miguel Ángel Martínez Prieto got in September 2007 their DEA under our supervision. At this moment, one student, Daniel
Aréchiga, is working on the dissertation for DEA and another student, Javier David Fernández García, has begun his doctoral studies and has applied for a doctoral grant (FPU).

3.2.2 Employees

Javier David Fernández García has been hired as técnico medio (April 2008–December 2008).

3.2.3 Publications

Articles in international journals


Lecture Notes in Computer Science


Communications in proceedings of international conferences


Communications in proceedings of national conferences


Chapters in books


3.2.4 International cooperation

The group has continued with the collaboration with Gonzalo Navarro from the University of Chile. As a result, Miguel Ángel Martínez made a stay there, and several common papers were published [4, 3, 5].

3.2.5 Technology transfer

Our cooperation with Telefónica Investigación y Desarrollo (TID) has produced a prototype, MyMoSe, of a Mobile Web Search system in the Morfeo open-source project. Some results of this cooperation have been presented or accepted at several international conferences (as explained above). This collaboration allowed the group to participate as partner in the Avanza TSI-020302-2008-25 project.

At this moment, we are also developing a new version of the BHSC (Biblioteca Histórica de Santa Cruz) with a model based on services close to DELOS proposal.

3.3 Subproject 3: indexing and compressing

3.3.1 Trainees


Three students got their DEA under our supervision: Óscar Pedreira Fernández, Diego Seco Naveiras, and Eloy Vázquez Fontenla. A high number (17) of Master Theses were supervised in the framework of this project. One was related to compression, 8 to indexing and 8 to digital libraries.

Within this project we implemented and distributed among students different small compressor modules useful to learn about this research area.

3.3.2 Employees

Two people have been hired técnico superior: Eloy Vázquez Fontenla (April 2007–) and Eduardo Rodríguez López (April 2007–July 2007).
3.3.3 Publications

Articles in international journals (JCR)


Cited by:


Communications in proceedings of international indexed conferences


Cited by:


Cited by:


Cited by:

Other international journals


Cited by:


Other international and national conferences


3.3.4 International cooperation

A researcher in the subproject is a highly reputed scientist, Gonzalo Navarro from the Universidad de Chile. In this period, the cooperation was reinforced and new links with international research groups were established.

Recently, collaboration with Gabriella Pasi (University of Milan) started. This relation will foster our participation in future European projects.

The impact of our research allowed Nieves R. Brisaboa, head of this subproject, to be invited to participate in the committee evaluating Shane Culpepper’s PhD Thesis, supervised by the Australian researcher Alistair Moffat, one of the best-known scientists in the compression field.

3.3.5 Technology transfer

We created a web site http://rosalia.dc.fi.udc.es/codes were all our compressors are freely distributed under the General Public License (GPL).

Our compression technology for cartography will probably be useful to build Geographic Information Systems. We expect software to be released soon.

References


