Received 28 October 2008
Revised 14 January 2009
Accepted 21 January 2009

An integrated system for managing the Andalusian Parliament's digital library

Luis M. de Campos, Juan M. Fernández-Luna,
Juan F. Huete and Carlos J. Martín-Dancausa
Departamento de Ciencias de la Computación e Inteligencia Artificial, ETSI Informática y de Telecomunicación, Universidad de Granada, Granada, Spain, and

Antonio Tagua-Jiménez and Carmen Tur-Vigil
Parlamento de Andalucía, Sevilla, Spain

Abstract

Purpose – The purpose of this paper is to present an overview of the reorganisation of the Andalusian Parliament's digital library to improve the electronic representation and access of its official corpus by taking advantage of a document's internal organisation. Video recordings of the parliamentary sessions have also been integrated with their corresponding textual transcriptions.

Design/methodology/approach – After analysing the state of the Andalusian Parliament's digital library and determining the aspects that could be improved both in the repository and access mechanisms, this paper describes each component of the developed integrated information system.

Findings – A methodology has been developed to tackle the problem and this could be applied to other similar institutions and organisations. Exploiting the internal structure of the parliament's official documents has also proved to be extremely interesting for users as they are directed towards the most relevant parts of the documents.

Originality/value – The paper presents an application of an information retrieval system for structured documents to a real framework and the integration of multimedia sources (e.g. text and video) for retrieval purposes.

Keywords Digital libraries, Information retrieval, Extensible Markup Language, Video, Parliament, Spain

Paper type Case study

1. Introduction and motivation

With the development of computers and the internet, most organisations working with documents (in every sense of the word), in a physical format, i.e. paper, witnessed the first computer revolution whereby such documents were converted to an electronic format. This allowed advantage to be taken of new technologies for managing document collections and resulted in a faster and easier user access. Many organisations and institutions, therefore, started to produce digitally formatted documents (e.g. in Portable

This work was jointly supported by the "Consejería de Innovación, Ciencia y Empresa de la Junta de Andalucía" (TIC-276), the Spanish "Ministerio de Educación y Ciencia" (TIN2005-02516, TIN2008-06566-C04-01), and the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

Program: electronic library and information systems
Vol. 43 No. 2, 2009
pp. 156-174

0033-0337
DOI 10.1108/00330330910954370

Document Format or PDF) which were then made available to a wider public by means of the web, and search engines with basic functionalities were incorporated to enable users to search and access these digital libraries.

A step forward could be taken, with the internet still at the centre but surrounded by a whole "gamut" of new technologies that work simultaneously with several types of media. Information access is now:

- *Faster* – because of the improvement in the underlying communication technology.
- *More accurate* – as the user obtains information which is closer to his or her information requirements.
- *Easier* – because less effort is required from the user.
- *More diverse* – the sources are not limited to text, a mixture of text, images, audio and video can now be incorporated daily to enrich an organisation's digital library.

The management of digital libraries and the digital libraries themselves must therefore evolve towards a more advanced digital framework, and the institutions must take up the new technological challenge in order to remain up-to-date.

Such is the case of the Parliament of Andalusia, the Southern Spanish autonomous region, which was established in 1982. One of the main objectives of democracy is that citizens are informed of any decisions made by their parliamentary representatives at any given time. National and regional governments are consequently obliged to inform the public of any work undertaken by parliament so that all the matters discussed are public knowledge. Initially, the Andalusian Parliament published a record of parliamentary proceedings which consisted of documents printed on paper with exact transcriptions of all the speeches by Members of Parliament (MPs) relating to every matter discussed. These bulletins were then sent to official organisations and public libraries so that they were publicly available.

The first challenge for the Andalusian Parliament in terms of technology was to create a digital library with all the records of parliamentary proceedings which users could easily access. They then changed their *modus operandi* and PDF documents were generated from the transcriptions. This file type is currently the most widely used in organisations for storing and spreading textual information as it enables electronic documents to be exchanged and viewed reliably and easily, regardless of the environment in which they were created.

Having prepared the e-documents, the organisation included a search engine on its web site (www.parlamentodeandalucia.es) so that any internet user (not just politicians or parliament employees) could use a form to submit a natural language or database query to obtain any relevant documents.

As an alternative to the manual transcriptions of the speeches, video recordings of the sessions were introduced to supplement the parliament's available information resources. Up until now, these two completely different media forms from the same source were dealt with separately: the user was either looking for text or a video but not both, intersynchronised with one query. This kind of multimedia retrieval is one example of a challenge posed by the technological revolution.

A second interesting example relates to the format of the documents in the digital library and (as mentioned) PDF was the chosen format. When a user submits a query, the full document is returned, and the information required might be found in a certain paragraph or section of the document in an MP's speech. The user must therefore read the entire document to find the required information and this wastes time. This is the classic perspective on information retrieval (IR). However, if the official parliamentary document was extremely well organised with a rich internal structure, we could take advantage of this and, rather than returning the full document, we could only return the part that best matches the user's information requirements. The retrieval system will then direct the user to specific parts of the document, which are relevant to their query and this saves time. Since PDF is not the most appropriate format to store a document's structure, it is therefore advisable to select another document format and for these purposes Extensible Markup Language (XML) is a very suitable option as it is able to capture the internal organisation of the textual materials. This therefore represents a move away from the classic field of IR to a new area of structured IR, which deals with explicitly organised documents, whereby the problem is not to retrieve a relevant document but rather relevant parts of it.

In this paper, we present a case study for the Andalusian Parliament and show the methodology that was designed and applied in the framework of a research project to improve the parliament's digital library in terms of its document collection and document access infrastructure (although there is no commonly agreed definition, these are the basic components of a digital library according to Seadle and Greifeneder (2007)). This methodology was of course particularised to the specific problem in hand but the steps could easily be generalised and adopted by any similar organisation.

We therefore present an integrated information system comprising various software modules which have been designed and implemented to improve access to the documents and videos in the Andalusian Parliament's repository by exploiting the documents' internal structure for retrieval purposes.

In the following section of this paper, we provide some background information about the Andalusian Parliament and its publications. Section 3 analyses the possible weaknesses detected in the digital library and possible solutions for strengthening it. Section 4 describes the methodology that we have designed for our case study and translation to a computer application. Section 5 introduces all the software modules (PDF-to-XML converter, video segmentation and synchronisation tools, search engine and its user interface). The final section outlines our conclusions and discusses various proposals for future work.

2. The Andalusian Parliament and its digital library

This section presents a brief introduction to the Andalusian Parliament and its official publications in order to contextualise this paper and introduce specific terminology.

As mentioned, the Andalusian Parliament was established in 1982 and to date there have been eight legislatures (periods of political activity of up to four years). The parliament edits two main official publications: the *Record of Parliamentary Proceedings* and the *Official Bulletin*. The first publication contains full transcriptions of all the MPs' speeches in each parliamentary session in which laws are passed or in the informative sessions held with MPs. Additional information such as the official agenda (the matters to be discussed as agreed by all the political groups, i.e. the

minutes of the meetings), results of possible votes, agreements, etc. is also included. In the second publication, the parliament publishes any texts and documents to be made available to the public about laws passed or to be processed.

There are three main types of sessions:

(1) *Plenary sessions.* These are attended by all MPs to debate an initiative.

(2) *Committee sessions.* These are attended by MPs according to different areas of interest (agriculture, economy, education, etc.) to discuss relevant initiatives.

(3) *Permanent parliamentary sessions.* These are attended by various duty MPs when parliament is not in session.

Parliament works around the concept of "parliamentary initiative", whereby an action taken by an MP or political party is discussed in a plenary or specific area committee session. These initiatives are included in the plenary or committee sessions and identified by means of an initiative code. Before a plenary session is held, the political parties represented in the house decide on the agenda for the session and this consists of a sequence of initiatives. These are subsequently grouped according to type and are previously published in the *Official Bulletin*. Once the agenda has been agreed, the speaker leads discussion of each point, allowing MPs to speak on and to discuss the corresponding initiative.

All official documents are published electronically on the Andalusian Parliament's web site, and PDF is currently the format most widely used by organisations (including the Andalusian Parliament) for storing and publishing textual information.

Once the PDF documents have been published, the Andalusian Parliament also provides a search engine so that users can consult the legislative collection by means of a database-like query using a web form. As the sessions are also video-ed, the parliament's digital library, with its records of parliamentary proceedings, is supplemented with videos on the web site which, while not able to be searched, can be browsed by date and viewed.

Since it was established 26 years ago, 5,975 documents have been published with the records of parliamentary proceedings and the official bulletins. While the size of this digital library is not excessive, it is a respectable figure and is constantly growing.

To help the reader understand the rest of the paper, we will now describe how the records of the parliamentary proceedings and the official bulletins are organised. The records of parliamentary proceedings comprise four distinct yet well-defined parts:

(1) *General information section*. This contains general information about the parliamentary session in question (e.g. type of session, legislature, date and presidency).

(2) *Agenda*. A list of the initiatives grouped according to type. This is decided by the political parties before the session and follows the format of initiative code, subject and proposer. For example, in the type "oral questions", all the questions are listed specifying for each one the question and the asker.

(3) *Summary*. A detailed description of the agenda which is created once the session has finished. Once again, all the initiatives are grouped according to type, and in addition to a description of the initiative (code, matter and proposer) they also include a list of MPs participating in the debate and the result of any vote. New agenda items may be added or others removed.

(4) *Development of the session*. For each point included in the summary section (following the initiative information), transcriptions of all the speeches are included and set out like a script for a play or a film.

The *Official Bulletin* is organised in the following way:

- *Summary*. In addition to identification and date of the issue, this includes a brief reference to each parliamentary initiative developed in the main body of the document and serves as a kind of table of contents. There exists a well-defined and static hierarchical taxonomy of initiatives designed by the parliament (e.g. law projects, oral and written questions, motions, etc.), so all the initiatives presented in each document are arranged in the corresponding place in the hierarchy.
- *Body of the bulletin*. Following the taxonomy presented in the table of contents, this develops the content of each initiative and includes information such as the initiative number, proposer, answerer (in the case of a question to parliament, for example), date, and the text itself explaining the request or answer.

So that these documents can be accessed, the institution's web team maintains a search engine whereby users search one type of publication at a time, expressing a query in natural language as well as specifying the date of publication, identification of the document and/or legislature (or ranges for this data) in a database-type query. The query returns a list of PDF files sorted according to publication date rather than relevance to the query.

It is worth mentioning that these two types of documents present a very rich, well-defined internal structure, which has not until now been exploited for retrieval purposes.

With respect to the users of the digital library, we could classify them in three groups:

(1) MPs;

(2) administrative workers of the parliament; and

(3) general public.

Although the users are relatively different, their needs are very similar, as they usually wish to read documents or parts of them which are relevant to a specific matter.

3. Detected weaknesses and possible improvements

In technological terms, the parliament's digital library (with its collection of documents and access methods) has certain weaknesses that need to be resolved in order to improve quality:

- As documents are searched in a database-type style, they are displayed in descending order of date and cannot therefore be shown in terms of their relevance to the user's query (those most relevant, first).
- In order to find relevant information, users must select the type of document they want to search. If they want to search various types of documents, they must submit identical queries for each document type.

- Documents are treated as a whole, i.e. as atomic units. Once a PDF file is retrieved, the user must search for the relevant information within the document, with the time-wasting that this involves for the user.
- Videos are only searched for by date; users must know the exact date of the session recording to locate the video they want to find.
- Users must watch the full video to find the relevant part with the information they require.
- There is no connection between the records of parliamentary proceedings and the videos. Although these two different types of media represent the same session, they are not interconnected and so there is no direct access between a retrieved PDF file and its associated video.

Bearing these weaknesses in mind, and considering leading technologies, we believe that improvements in digital library management and use are possible, and therefore propose the following points:

- By applying IR methods and techniques, and given a query formulated in natural language, it would be possible for users to obtain a ranked list of all types of documents which are sorted according to relevance to the query (the higher the document in the ranking, the more relevant it is). It is therefore only necessary for users to read the uppermost-ranked documents to find the information they require.
- Thanks to the document's internal organisation, it is possible to apply structured IR techniques so that the search engine could return a ranked list of document parts (rather than full documents) in response to a query. This is clearly an important advantage because the system directs the user to the exact text unit where the relevant information is to be found regardless of its size and it is not necessary for the user to waste time finding what they are looking for. The main requirement of this new approach is to represent the documents in a much more flexible format, where their structure is explicit and easily managed. XML is the most suitable option for such a task and this meta-language has been used extensively in digital libraries (Kim and Choi, 2000; Yeates, 2002; Chang, 2005). Additionally, the search engine should be endowed with specifically designed retrieval models for managing structured documents and retrieving parts of them.
- A powerful query mechanism that would enable the user to formulate more accurate queries to help them find the information they require. In addition to natural-language queries (with possible restrictions on the type of document, legislature, date, etc.), the user could specify what to look for (content), where to search, and the output required. In the context of structured IR, this is called a content and structure query (CAS queries) and offers the possibility of specifying which parts of the documents to search and which to return.
- A link between the records of parliamentary proceedings and their associated videos, so that users could read the most relevant parts of the documents and watch the section of the video corresponding to the relevant part of text. Not only do the users obtain various supplementary multimedia sources in response to a single query, but they are also directed to the relevant part of the video so that they do not waste time trying to find it themselves in the full video. In order to establish such an association, additional tools must be developed and these

include an application to segment videos and an annotation tool to synchronise text and video. We should highlight that with this connection, videos are retrieved and attached to a portion of text as a result of a query, unlike other specific and more complex video retrieval techniques (Petrelli and Auld, 2008).

By developing an information system for the parliament's digital library, we intend to solve some of the weaknesses detected, and to facilitate internal processes and access to the official collections, making these more efficient and effective.

4. Methodology and system architecture

In this section, we will describe the methodology that we have designed and followed to update technologically the Andalusian Parliament's digital library. Although this methodology has been particularised to the specific problem in question, it could easily be generalised and exported so that it may be applied and adapted to any other institution or organisation with similar requirements. Our final aim is to develop an information system that covers the whole process of edition, publication and access of the elements comprising the digital library.

In addition to explaining the methodology, and as a consequence of its application, we will also show the software infrastructure of the general information system in terms of the required modules and relationships between them.

But before this description, we have to mention that this project has been carried out by a team composed of four members from the University of Granada, three researchers and a full-time research fellow, and six collaborators from the Andalusian Parliament. The duration of the project has been three years.

The following steps should be taken in order to satisfy the general requirements presented in the previous section:

(1) Design of a Document Type Definition (DTD) to represent the internal structure of the records of parliamentary proceedings and official bulletins. The two types of official documents must be analysed to extract the exact internal structure governing them. This is a manual process which results in a DTD file which contains an explicit description of the document's structure.

(2) Document conversion from PDF to XML. As the original documents are in PDF format which does not usually store information about a document's logical organisation and since the IR system (IRS) will work explicitly with the structure and content by means of XML documents, it is necessary to convert the formats automatically.

(3) Video segmentation and synchronisation with the text. In order to link a part of the record of parliamentary proceedings with its corresponding part of the video (for example, the text of an MP's speech and the speech itself in the video), it is necessary to split the video into segments. After this semi-automatic segmentation process, the parts of the text and the segments for the same speech must be associated (synchronised). This is a mainly manual but computer-assisted process.

(4) Design of a query mechanism. The way users express their information requirement is very important. The different alternatives (natural language, CAS, restrictions on legislature, dates, type of document, etc.) call for the design of a complete, intuitive mechanism to facilitate this task.

(5) Design of a retrieval model for structured documents and development of the corresponding IRS.

(6) Preparation of the document collection (indexing). In order to access the document collection by means of queries, the IRS must adapt the XML documents to its internal data structures, so that they may be effectively and efficiently accessed.

(7) Design and development of a graphic user interface and integration of the search engine.

The application of these steps results in the software modules shown in Figure 1, which shows the different application elements and the relationships between them in terms of inputs and outputs.

Beginning with the PDF collection, one of the first modules to be applied is the PDF-to-XML converter. This will transform all the PDF documents to XML format by extracting the text and placing it in the correct position in the structure of the record of parliamentary proceedings and the official bulletin. In parallel, each video of the parliamentary proceedings will be automatically segmented by the video segmentation tool to obtain a series of segments marked by a start and end time. These video portions could be manually edited to adjust the original output given by the software.

Once both processes have been completed, the records of parliamentary proceedings and their recordings must be synchronised. In this case, given an XML record of parliamentary proceedings, time attributes pointing to the video segments are included in the XML tags associated with the speeches. A time propagation is then performed

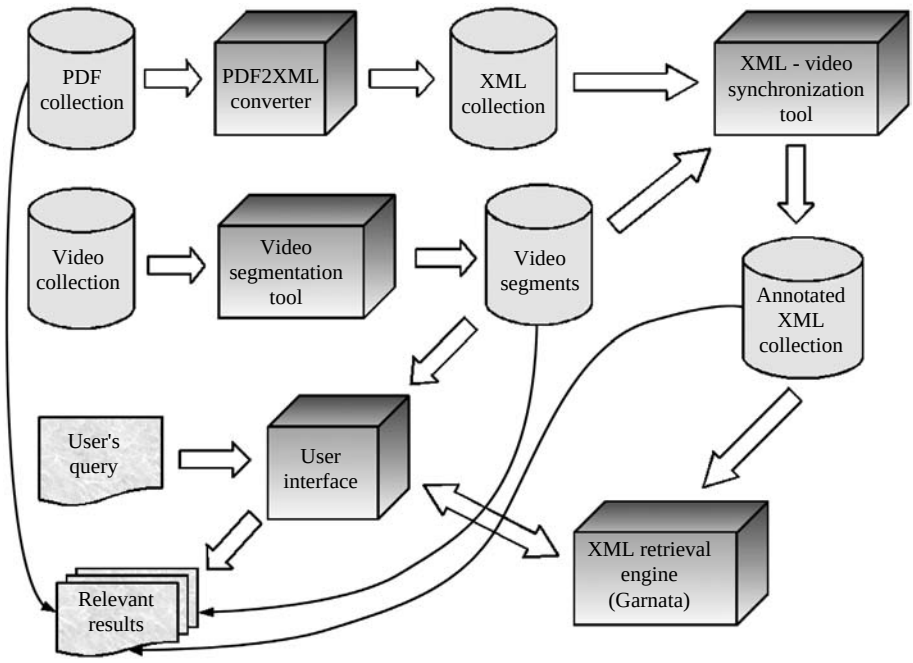

Figure 1.
Architecture of the system

towards the upper elements of the document structure. At this moment, and once the search engine has indexed the XML collection so that it may be managed efficiently, the user is able to formulate a query on the web-based user interface. This query is passed to the search engine which then computes the most relevant XML elements using a retrieval model based on Bayesian networks and influence diagrams and implemented in the Garnata IRS. These elements are ranked in decreasing order according to relevance and shown to the user, and can be arranged in various different ways (e.g. all the results, grouped according to document, or the best entry point for each relevant document). Users can then view the XMLs, PDFs and videos to find the information they are looking for.

5. Description of system components

This section will explain all the components shown in Figure 1 and those mentioned in the previous section in more detail.

5.1 Creating the XML collection: converting PDF documents to XML

As mentioned, the Andalusian Parliament's legislative collection could be exploited from a structured IR perspective. To achieve this objective, the first problem to resolve is how to obtain the internal document structure from a PDF in order to feed the structured IRS. More specifically, not only is it necessary to mine the text to search for the structure but we must also attach the text found in the document to the corresponding part of the structure. The result of this process is a file containing both elements (content and structure) in XML format.

Although there are various publicly available PDF-to-XML converters, the problem common to most of these is that the extracted structure is not usually the document's logical organisation but rather a layout-related document. As a result, the XML file obtained from a PDF document is organised according to page, paragraph, footnotes, etc. As this option is clearly not appropriate for our purposes, we must write our own converter. Research work on this subject can be found in specialist publications (Déjean and Meunier, 2006; Gurcan *et al.*, 2003), and while both of these approaches work directly with the internal information included in the PDF file, they do not capture the logical organisation of the documents, as we require.

In order to extract the text contained in the files, we therefore have two options: either to use an intermediate tool to obtain a text file which will be the input of the XML converter, or to extract the text by means of a toolkit that allows the PDF to be accessed but which is totally integrated in the developed software. We have opted for the first choice and used an external PDF-to-text converter. More specifically, we have used the open source and command line utility *pdftotext* which is integrated in the XPDF package (www.foolabs.com/xpdf/home.html), and the output of which will be the input of our XML converter. Our application then generates the final XML file starting from the text files with the textual content of the PDF. The software has been developed in Java (it is a command line application) and performs the following steps (Fernández-Luna *et al.*, 2008).

Using a lexical analyser, the content of these text input files is processed to produce a sequence of symbols called tokens as the output. The programme automatically introduces a group of structural components or labels in the text which correspond to the tokens, indicating the bounds of the different sections found in the document

(e.g. types of subjects discussed, initiative code, MPs' speeches, etc.). Another task performed by the lexical analyser is to eliminate noise, i.e. the deletion of those parts of the document which are not important from a semantic point of view.

When the lexical analyser has finished, the converter runs a syntax analyser developed in JavaCC to generate a grammar to detect the tokens mentioned in the previous step. To create XML files, we use the Document Object Model (DOM) methodology based on the building of a tree in memory with XML tags for nodes. When a token is detected, a new node or a group of nodes is therefore created in the DOM tree, generating all the hierarchical structure of the XML format. The tree is subsequently transformed into an XML file, and is finally validated in order to assure a valid and well-formed file.

5.2 Dealing with videos: segmentation and annotation

In terms of the session videos, as the main objective is to enable the user to access not only the most appropriate unit of text but also the video segment associated to that text, all the elements of the XML documents corresponding to speeches must be synchronised with their corresponding video segments. In order to achieve this, a previous step is to segment the videos. In the case of the Andalusian Parliament, there are only four cameras recording the sessions and so recording is quite simple, and the segments will therefore coincide with camera changes. A segmentation module will perform this task and produce the segments and their key frames (segmentation obtained automatically can be edited manually). The next step is to synchronise text and video. Using an annotation tool, an expert user will proceed to associate visually each segment with the corresponding XML tag containing the transcription of the audio of the video segment. The output of this process is the XML of a session with time stamps to indicate where the corresponding speech starts and ends in the video.

Our objective is to find a method which is capable of segmenting the videos of the Andalusian Parliament correctly and efficiently and where the complexity of implementation remains as low as possible (for further information about segmentation (Koprinska and Carrato, 2001; Camastra and Vinciarelli, 2007). These videos contain long scenes, with few movements and sudden changes. Four fixed cameras are installed in the Andalusian Parliament: one films the speaker, another shows a general view of the house, and two show the MPs. The video sequences are therefore usually very static and it is extremely easy to detect camera changes. If we consider segmentation using shot differences (Cotsaces *et al.*, 2005; Lienhart, 1999), we can see that differences are larger when there are camera changes, and smaller in the same segment because the cameras are static and there is hardly any movement. We shall therefore use this method (considering colour to be the feature on which the method is based), basically because of the method's simplicity and good results. As no sophisticated method is needed because of the features of the videos, we decided to implement our own segmentation algorithm which could be adjusted to them. The designed segmentation algorithm is therefore based on detecting differences between shots, and more specifically, differences between the grey tones of the shots (de Campos *et al.*, 2008a, b).

The basis for this algorithm is a comparison of the histograms by computing a distance measure of two consecutive shots. If this value is greater than a certain threshold, then there is a change in shots (a series of shots are considered to be included in the same segment if the difference between their histograms is low).

The initial algorithm may be improved as follows in order to increase the efficiency of the process and reduce the number of mistakes:

- Apply a convolution filter to avoid errors when the shots belong to the same segment.
- Decrease the number of histogram comparisons performed by discarding a number of shots between each studied pair (rather than analysing each consecutive pair).
- Compare shots only with a significant part of the image, instead of the whole image.

With these improvements, the algorithm is fast and accurate as it is totally tuned and adapted to the particular characteristics of the session videos.

Once a video has been automatically segmented, the software offers the possibility of editing the segmentation manually. The output of this process is a set of segments, which are represented by a key frame. The user is therefore allowed to edit the segments, combining them if they are contiguous, dividing segments in two or just playing them. Figure 2 shows a screen shot of this tool.

The synchronisation stage is performed with a second application. The input of this process will be the sets of segments found in the video corresponding to a parliamentary session and the transcription of the speeches given in the House for that video in XML format. The output of the process will be the same XML document containing the transcription synchronised with the video by means of time stamps in the elements of the document. The annotation tool will consist of the manual association of segments with the corresponding elements in the XML document, so each tag will have a link to its corresponding part of the video.

Figure 3 shows the user interface of the annotation tool. In the annotation process, the user first selects a segment in the video, then finds the node in the XML document

Figure 2.
Screen shot of the segmentation edition tool

Figure 3.
User interface of the synchronization tool

containing the audio transcription of that segment, and the former is finally associated with the latter by means of a drag-and-drop action. These steps are repeated until all the segments have been assigned to a node in the document.

With the association of a segment to an XML element in the document, we introduce a pair of attributes to the corresponding tags and these contain the start and end times of the segment. This information will be enough to access the video segment in retrieval time.

The segmentation algorithm and the segmentation and synchronisation tools are described in more detail in de Campos *et al.* (2008a, b).

5.3 Garnata, the search engine: indexing and retrieval

Standard IRSs (Manning *et al.*, 2008) treat documents as atomic entities, so usually only entire documents constitute retrievable units. More elaborate document representation formalisms (e.g. XML) allow us to represent so-called structured documents, the content of which is organised around a well-defined structure enabling the semantics of long, complex documents to be described (Chiaramella, 2001) and examples of these documents are books, textbooks, scientific articles, etc. and also parliamentary official bulletins and records of proceedings. Structured IRSs view documents as aggregates of interrelated structural elements that need to be indexed, retrieved, and presented both as a whole and separately according to user requirements. In other words, given a query, a structured IRS must retrieve the set of document components that are most relevant to the query rather than full documents. While the document structure is "flattened" and not exploited by classical retrieval methods, structured IR models exploit the content and structure of documents to estimate the relevance of the document components to queries, usually based on the aggregation of the estimated relevance of their related parts (Lalmas and Ruthven, 1998). This is clearly the most appropriate approach for exploiting the advantages of the implicit structure of the Andalusian Parliament's official collections so that more accurate and relevant material may be offered.

The search engine for retrieving relevant material is Garnata (de Campos *et al.*, 2006), an IRS specially designed to work with structured XML documents. It is responsible for indexing the XML version of the official documents, with the possibility of stop-word removal and stemming (in this case in Spanish), and retrieving them given a query. The retrieval engine has been developed upon a retrieval model for XML documents based on probabilistic graphical models (de Campos *et al.*, 2004, 2005).

Garnata computes the relevance degree of each component or structural unit in a document by combining two different types of information:

(1) *The specificity of the component in relation to the query* – the more terms in the component that appear in the query, the more relevant the component is, i.e. the more clearly the component (or at least part of it) is only about the topic in the query.

(2) *The exhaustivity of the component in relation to the query* – the more terms in the query that match terms in the component, the more relevant the component is, i.e. the more clearly the component is about the topic in the query.

The components that best satisfy the user's information requirements expressed in the query should be, simultaneously, as specific and exhaustive as possible.

These two dimensions concerning a component's relevance to the query are calculated in different ways. In order to compute the specificity, the probability of each component's relevance is obtained through an inference process in a Bayesian network representing the structured document collection, where the nodes are the different terms and document components, and the topology reflects the inclusion relationships between the document components together with the membership of each term to the corresponding document components (Figure 4). The exhaustivity is obtained by first defining the utility of each document component as a function of the proportion of the terms in the query that appear in this component. The Bayesian network is then transformed into an influence diagram (by adding the corresponding decision and utility nodes) which computes the expected utility of each component by combining the

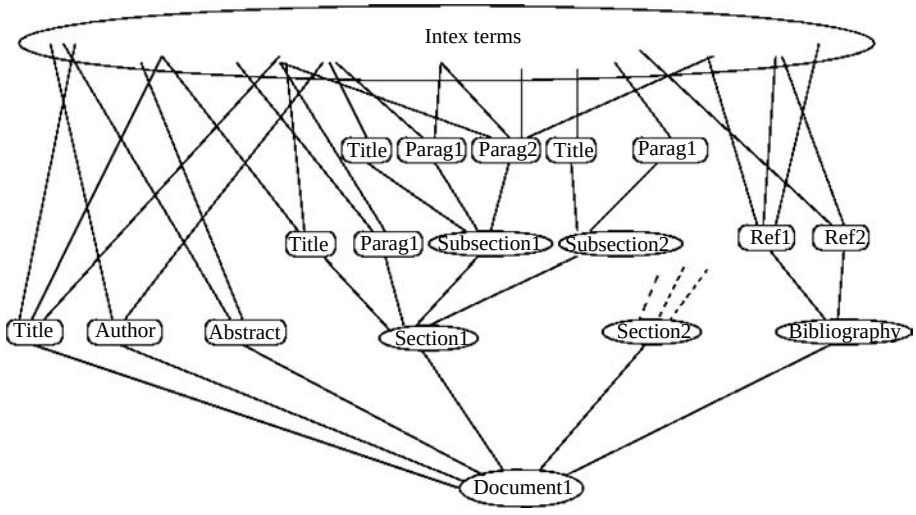

Figure 4.
An example of Bayesian network representing a structured document

relevance probabilities and the utilities in a principled way. Once this type of relevance score has been computed for each unit, the next operation is to return a ranking of units, i.e. a list of structural elements, with different granularity, sorted in decreasing order of relevance.

In terms of the query, the system is able to perform inferences in the underlying model to solve two types: content queries and the more elaborate content and structure queries. The former are natural-language sentences which are transformed into bag-of-word queries; the latter are queries in which the users specify structural restrictions in terms of the units in which to search the query, and the units they are interested in as a result of the process. The Narrowed Extended X Path I (NEXI) language (Trotman and Sigurbjörnsson, 2005) is the underlying query language used to represent this type of query. In order to manage content and structure queries, the Garnata-supported retrieval model has been provided with new capabilities, which basically allow the NEXI queries to be parsed, breaking them down into their corresponding sub-queries, retrieving components for each sub-query and combining the results into a single output, taking into account the structural restrictions imposed by the original query.

In order to determine whether the model's method of retrieval is effective enough to be used in the setting of the Andalusian Parliament, the model's performance was evaluated by our participating in various editions of the initiative for the evaluation of XML retrieval (INEX) (http://inex.is.informatik.uni-duisburg.de and www.inex.otago.ac.nz/) (Fuhr *et al.*, 2008) for an example of the 2007 workshop). Acceptable results were obtained in comparison with other participants in the different tasks on which we tested our system (de Campos *et al.*, 2007, 2008a, b). With Garnata as the search engine in the integrated system presented in this paper, we therefore ensure that a high quality is achieved in the retrieval task. We must also highlight the fact that the system is efficient in terms of retrieval as it usually only takes a few seconds to return the relevant material.

5.4 The user interface

The search engine is accessible through a web application which follows the client/server paradigm. Its user interface has been carefully designed in collaboration with staff at the Andalusian Parliament. The objective was to develop a simple and intuitive interface.

A user can access information in the Andalusian Parliament's legislative collection by means of two forms on the parliament's web page http://irutai.ugr.es/WebParlamento. The first is for content queries (Figure 5) where the query is written in a text field and could be restricted by specifying the legislature number, kind of document (records of parliamentary proceedings or official bulletins), publishing dates, or range of documents. The CAS query is also formulated using a form, which helps the user specify structural restrictions to prevent the end-user knowing the NEXI language in which this kind of XML query is formulated. The results of a search are shown in Figure 6.

It is also possible to indicate how the results are arranged, so the application helps the users find the units satisfying their information requirements:

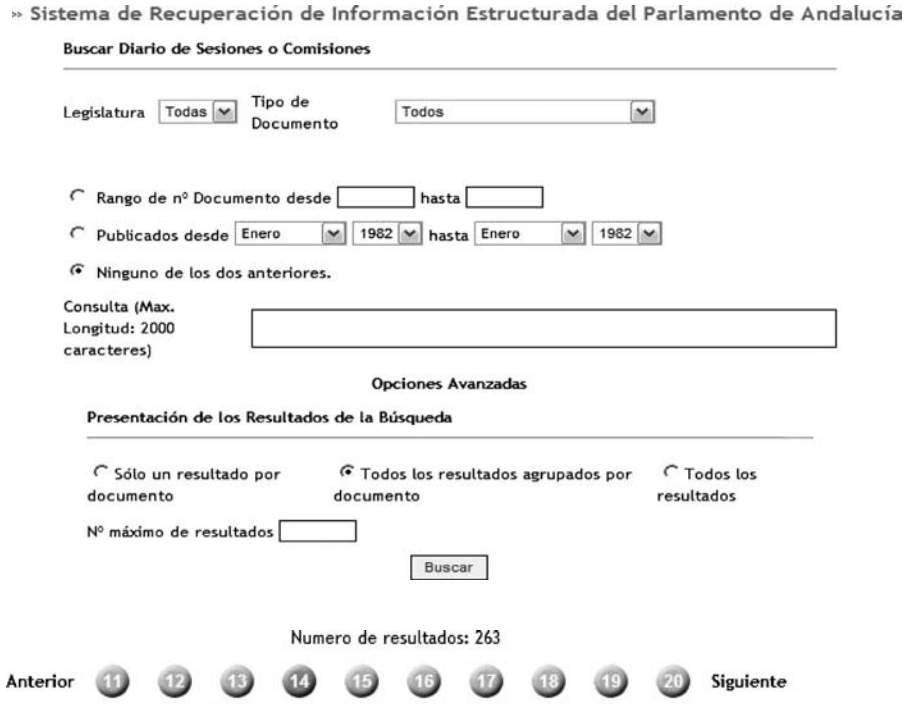

Figure 5.
Query form

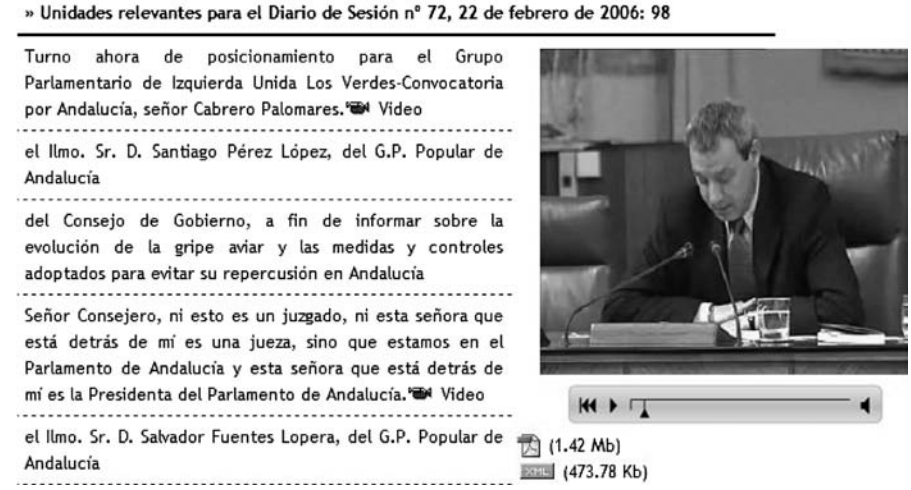

Figure 6.
Results of a search

- Only one result per document: the system will show only one result per document. This single document part should correspond to the best entry point for starting to read the relevant text in the document.

- All the results grouped by document: for each document, the search engine will return every relevant unit according to its relevance.
- All the results: all the relevant units (without any association) are presented to the user in decreasing order of relevance.

Once the search engine has computed the relevance of the structural units in the collection, the results are presented on a second web page in groups of ten, sorted decreasingly in terms of relevance to the query. For each result, a brief portion of the text of the structural unit is provided together with a link to the corresponding PDF document containing this unit and a link to the XML document visualised in Hyper Text Markup Language format, highlighting the relevant units. If there is a video for a unit, then there will also be a link to this video so that the user will be able to watch the video segment corresponding to this structural unit. In Figure 6, we show an example of how the results are presented by playing a video segment when the "All the results grouped by document" option is selected.

The video player is implemented using Flash technology as it allows videos to be played on any platform. The format of these videos is lighter than their original format, saving broadband in the video transmission.

6. Conclusions and future work

In this paper, we presented our experience of improving the digital library of the official collections generated by the Andalusian Parliament.

In the first step, we proposed a new way of electronically representing the records of parliamentary proceedings and official bulletins in order explicitly to represent the internal structure of these documents. By using XML to specify a document's internal organisation, and with the aid of the IR, our presented search engine can control the granularity of the retrieved elements according to their relevance in relation to a query.

The original documents are stored in PDF format, which is suitable for representing content and the external (but not internal) structure and so it is necessary to convert the documents. Our converter is able to extract automatically the text of each document and place it in the correct position of the hierarchical structure of an XML document.

Not only does the parliament's digital library contain texts but videos also play an important role. As there are two representations of the same information (a record of parliamentary proceedings with the transcriptions of the speeches and their corresponding recordings), it seems logical to access both at the same time under the same query. Two previous steps are required: first, the videos must be segmented into parts with continuity; and second, the XML files must be synchronised with these segments so that when a structural unit is retrieved, it will also be possible to watch the associated video.

The paper also presented Garnata – a search engine for XML documents based on Bayesian networks. This engine is effective and efficient and its user interface is thorough and easy to use. The results can therefore be displayed in various ways, enabling the user to consult documents in XML or PDF format and watch the session videos.

Because this integrated tool is richer and more flexible, the Andalusian Parliament's official collection may be easily accessed and document representation improved. To the best of our knowledge, and although IR techniques have been widely applied to

digital libraries (Chowdhury and Chowdhury, 2000), the experience presented in this paper is one of the first to exploit the internal structure of documents in a digital library from the point of view of retrieval in a real environment.

One of the main conclusions of this work is that all the steps and developments for this project could easily, and with relatively little effort, be exported to other similar institutions or organisations. The only module that would require a major modification would be the PDF to XML converter as it should be adapted to the structure of the new collection. The rest of the application would suffer minor changes and could be used almost directly.

Great efforts were made in the PDF-to-XML conversion and video segmentation and synchronisation stages. In terms of the former, it took us a long time to develop the converter and perform the conversion; with respect to the latter, although the segmentation is an automatic process, which can be refined manually, synchronisation is mainly manual, and again this step usually takes time. As lessons learnt, we might mention the need to invest in alleviating the workload of these stages in order to achieve greater processed inputs to reduce the processing time.

One important outcome of this project was that the parliamentary department in charge of publishing parliamentary records of proceedings and official bulletins has changed the way it works. Instead of producing documents in Microsoft Word and later in PDF format, XML files are being generated directly which can then be fed into the search engine without the need for conversion. In addition, with the novel video management system that they are adopting, they are able to segment new recordings at speech level and synchronise them with the XML documents easily and automatically. This will clearly lead to an improvement in how the corpus is dealt with.

Further work in the future could be carried out in order to improve document access:

- design of relevance feedback techniques for XML documents so that user opinions about the relevance of the units to the original query may be incorporated in order to obtain more relevant documents;
- development of personalisation techniques to include the user's special characteristics and features in the search process;
- incorporation of the user's context in the retrieval;
- integration of the Eurovoc thesaurus (http://europa.eu/eurovoc/) in the application to help the user formulate a query; and
- perform a deep user evaluation to test the usability of the retrieval module.

References

Camastra, F. and Vinciarelli, A. (2007), "Video segmentation and key frame extraction", *Machine Learning for Audio, Image and Video Analysis: Theory and Applications Advanced Information and Knowledge Processing*, Vol. XVI, Springer, Berlin, pp. 413-30.

Chang, N. (2005), "Data manipulation in an XML-based digital image library", *Program: Electronic Library and Information Systems*, Vol. 39 No. 1, pp. 62-72.

Chiaramella, Y. (2001), "Information retrieval and structured documents", *Lecture Notes in Computer Science*, Vol. 1980, pp. 291-314.

Chowdhury, G.G. and Chowdhury, S. (2000), "An overview of the information retrieval features of twenty digital libraries", *Program: Electronic Library and Information Systems*, Vol. 34 No. 4, pp. 341-73.

Cotsaces, C., Gavrielides, M. and Pitas, I. (2005), "A survey of recent work in video shot boundary detection", *Proceedings of 2005 Workshop on Audio-Visual Content and Information Visualization in Digital Libraries*, available at: http://poseidon.csd.auth.gr/papers/PUBLISHED/CONFERENCE/pdf/Cotsaces05a.pdf

de Campos, L.M., Fernández-Luna, J.M. and Huete, J.F. (2004), "Using context information in structured document retrieval: an approach using influence diagrams", *Information Processing & Management*, Vol. 40 No. 5, pp. 829-47.

de Campos, L.M., Fernández-Luna, J.M. and Huete, J.F. (2005), "Improving the context-based influence diagram for structured retrieval", *Lecture Notes in Computer Science*, Vol. 3408, pp. 215-29.

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F. and Martín-Dancausa, C.J. (2008a), "The Garnata Information Retrieval System at INEX'07", *Lecture Notes in Computer Science*, Vol. 4862, pp. 57-69.

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F. and Romero, A.E. (2006), "Garnata: an information retrieval system for structured documents based on probabilistic graphical models", *Proceedings of the Eleventh International Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Paris, France*, pp. 1024-31.

de Campos, L.M., Fernández-Luna, J.M., Huete, J.F., Martín-Dancausa, C.J. and Romero, A.E. (2007), "Influence diagrams and structured retrieval: Garnata implementing the SID and CID models at INEX'06", *Lecture Notes in Computer Science*, Vol. 4518, pp. 165-77.

de Campos, L.M., Fernández-Luna, J.M., García, J.M., Gómez, F., Huete, J.F. and Martín-Dancausa, C.J. (2008b), "A video segmentation and annotation tool for parliamentary recordings and transcriptions", *Proceedings of the IADIS International Conference Informatics*, pp. 35-42.

Déjean, H. and Meunier, J. (2006), "A system for converting PDF documents into structured XML format", *Lecture Notes in Computer Science*, Vol. 3872, pp. 129-40.

Fernández-Luna, J.M., Huete, J.F., Gómez, M. and Martín-Dancausa, C.J. (2008), "Development of the XML digital library from the Parliament of Andalucía for intelligent structured retrieval", *Lecture Notes in Artificial Intelligence*, Vol. 4994, pp. 417-23.

Fuhr, N., Kampas, J., Lalmas, M. and Trotman, A. (2008), "Focused access to XML documents", *Proceedings of the 6th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2007, Dagstuhl Castle, Germany*, Lecture Notes in Computer Science, No. 4862, Springer, Berlin, pp. 57-69.

Gurcan, A., Khramov, Y., Kroogman, A. and Mansfield, P. (2003), "Converting PDF to XML with publication-specific profiles", *Proceedings of the XML Conference and Exposition 2003, Philadelphia, PA*, available at: www.idealliance.org/papers/dx_xml03/index.html

Kim, H. and Choi, C. (2000), "XML: how it will be applied to digital library systems", *The Electronic Library*, Vol. 18 No. 3, pp. 183-9.

Koprinska, I. and Carrato, S. (2001), "Temporal video segmentation: a survey", *Signal Processing: Image Communication*, Vol. 16 No. 5, pp. 477-500.

Lalmas, M. and Ruthven, I. (1998), "Representing and retrieving structured documents with Dempster-Shafer's theory of evidence: modelling and evaluation", *Journal of Documentation*, Vol. 54 No. 5, pp. 529-65.

Lienhart, R. (1999), "Comparison of automatic shot boundary detection algorithms", *Proceedings of SPIE, Storage and Retrieval for Image and Video Databases VII*, Vol. 3656, pp. 290-301.

Manning, C.D., Raghavan, P. and Schütze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge.

Petrelli, D. and Auld, D. (2008), "An examination of automatic video retrieval technology on access to the contents of an historical video archive", *Program: Electronic Library and Information Systems*, Vol. 42 No. 2, pp. 115-36.

Seadle, M. and Greifeneder, E. (2007), "Defining a digital library", *Library Hi Tech*, Vol. 25 No. 2, pp. 169-73.

Trotman, A. and Sigurbjörnsson, B. (2005), "Narrowed extended XPath I (NEXI)", *Lecture Notes in Computer Science*, Vol. 3493, pp. 16-40.

Yeates, R. (2002), "An XML infrastructure for archives, libraries and museums: resource discovery in the COVAX project", *Program: Electronic Library and Information Systems*, Vol. 36 No. 2, pp. 72-88.

Corresponding author
Juan M. Fernández-Luna can be contacted at: jmfluna@decsai.ugr.es