DEFINING A WORKFLOW PROCESS FOR TEXTUAL AND GEOGRAPHIC INDEXING OF DOCUMENTS*

Nieves R. Brisaboa, Ana Cerdeira-Pena, Miguel R. Luaces, Diego Seco
Database Lab., University of A Coruña, Campus de Elviña, S/N 15071, A Coruña, Spain
{*brisaboa,acerdeira,luaces,dseco*}*@udc.es*

Keywords: GIR, workflow, index structure, ontology.

Abstract: Many public organizations are working on the construction of spatial data infrastructures (SDI) that will enable them to share their geographic information. However, not only geographic data are managed in these SDIs, and, in general, in Geographic Information Systems (GIS), but also many textual documents must be stored and retrieved (such as urban planning permissions and administrative files). Textual index structures must be integrated with GIS in order to provide an efficient access to these documents. Furthermore, many of these documents include geographic references within their texts. Therefore, queries with geographic scopes should be correctly answered by the index structure and the special characteristics of these geographic references, due to their spatial nature, should be taken into account.
We present in this paper a workflow process that allows a gradual and collaborative creation of a document repository. These documents can be efficiently retrieved using queries regarding their texts and regarding the geographic references included within them. Moreover, the index structure and the supported query types are briefly described.

1 INTRODUCTION

The research field of Geographic Information Systems (Worboys, 2004) has received much attention during the last years. Recent improvements in hardware have made the implementation of this type of systems affordable for many organizations. Furthermore, a cooperative effort including the definition of standards and specifications for interoperable systems, has been undertaken by two international organizations: ISO (ISO/IEC, 2002) and the Open Geospatial Consortium (Open GIS Consortium, Inc., 2003). This effort is making possible that many public organizations are working on the construction of spatial data infrastructures (SDI) (Global Spatial Data Infrastructure Association, 2008) that will enable them to share their geographic information. However, these geographic infrastructures manage not only geographical information but also textual information (such as urban planning permissions and administrative files). Therefore, textual index structures must be integrated in these infrastructures to provide an efficient access to these documents.

The research field of Information Retrieval (Baeza-Yates and Ribeiro-Neto, 1999) has been active for the last decades. The growing importance of Internet and the World Wide Web have made it one of the most important research fields nowadays. Many different index structures, compression techniques, and retrieval algorithms have been proposed in the last few years. More importantly, these proposals have been widely used in the implementation of document databases, digital libraries, and web search engines. Altough many of the documents stored in these digital libraries and documents

*This work has been partially supported by "Ministerio de Educación y Ciencia" (PGE y FEDER) ref. TIN2006-16071-C03-03, and by "Xunta de Galicia" ref. 2006/4 and ref. 08SIN008E.

databases include geographic references, these ones are rarely used in information retrieval systems.

During the last decades, these two research fields have advanced independently. Pure textual techniques focus only on the language aspects of the documents and pure spatial techniques focus only on the geographic aspects of the documents. None of them is suitable for a combined approach to information retrieval because each one completely neglects the other type of information. As a result, there is a lack of system architectures, index structures, and query languages that combine both types of information. Some recent proposals (Lieberman et al., 2007; Chen et al., 2006; Martins et al., 2005) define new index structures that take into account both the textual and the geographic aspects of a document. These proposals are the origin of a new research field called Geographic Information Retrieval (GIR).

In (Luaces et al., 2008), we present an architecture of a GIR system and an index structure that improve the query capabilities of other proposals. However, this architecture is not flexible enough to be used in organizations where the number of documents is constantly increasing. In public organizations (e.g. city councils), where new planning permissions or administrative files are generated every day, a workflow process must be implemented to define all the tasks for indexing a document in the repository. Moreover, there are some tasks that must be performed before the document indexing (e.g. metadata storage, scanning, OCR, etc.). These tasks were not taken into account in the presented architecture that assumes a static document collection.

Therefore, this paper proposes a set of strategies for the workflow management of the repository creation process and a general system architecture supporting them. The proposed strategies improve the performance of the system, ensuring that all the necessary tasks are correctly performed, and facilitating the work of the people devoted to this activity. In addition, textual, spatial, and hybrid queries (e.g. *planning permissions of civil buildings in A Coruña*) can be solved by means of the index structure integrated in the system.

The rest of the paper is organized as follows. Some related work is presented in the next section. Section 3 presents the general architecture for the workflow management in the digitalization process. Then, in the Section 4 we briefly describe the index structure and the supported query types. Finally, Section 5 presents our conclusions and future lines of work.

2 RELATED WORK

Inverted indexes are considered the classical text indexing technique (Baeza-Yates and Ribeiro-Neto, 1999). An inverted index associates to each word in the text a list of pointers to the positions where the word appears in the documents. The main drawback of these indexes is that geographic references are mostly ignored because place names are considered words just like the other ones. If the user poses a query such as *hotels in Spain*, the place name *Spain* is considered a word, and only those documents that contain exactly that word are retrieved.

Regarding indexing geographic information, many different spatial index structures have been proposed throughout the years. A good survey of these structures can be found in (Gaede and Günther, 1998). A drawback of spatial index structures is that they do not take into consideration the geographic ontology of the real world. Internal nodes in the structure are meaningless in the real world and it is not possible to associate location-specific information to these nodes because there is no relation at all between the nodes in the spatial index structure and real world locations.

Some work has been done to combine both types of indexes. The papers about the SPIRIT (Spatially-Aware Information Retrieval on the Internet) project (Jones et al., 2004; Vaid et al., 2005) are a very good starting point. Regarding our work in this research area, in (Luaces et al., 2008) we present an architecture of a GIR system and an index structure that combines an inverted index, a spatial index, and an ontology-based structure. Pure textual queries, pure spatial queries, and hybrid queries can be solved by this index structure that is described in Section 4.

Finally, regarding our work in document management systems and workflow processes, in (Places et al., 2007) we present a set of strategies to face the management of the workflow of the digital library building process and a general system architecture supporting them. The paper also presents a tool developed following that architecture. This tool provides an integrated environment where all tasks involved in the repository building can be performed. As we noted before, in this work we extend the architecture

to include new tasks that make the index able to solve queries taking into account the spatial nature of the geographic references included in the text of the documents.

3 SYSTEM ARCHITECTURE

According to (Hollingsworth, 1995), a workflow is concerned with the *automation of procedures where documents, information, or tasks are passed between participants following a defined set of rules to achieve or contribute to an overall business goal; the computerized facilitation or automation of a business process, in whole or part.* Workflow management systems can be classified in several types depending on the nature and characteristics of the process (van der Aalst and van Hee, 2002; Fischer, 2003). Collaborative workflow systems automate business processes where a group of people participate to achieve a common goal. This type of business processes involves a chain of activities where the documents, which hold the information, are processed and transformed until that goal is achieved. We based the architecture of the system in this model because the problematic of building a document repository fits perfectly in it.

In general, we can differentiate three user profiles involved in the repository building:

- *Administrator.* Administrators are responsible for the process as a whole. They are in charge of assigning tasks to different workers and controlling the state of each digitalized document.
- *Advanced users.* Advanced users are in charge of carrying out critical activities such as metadata storage or reviewing the geographic references extracted from the texts obtained by the OCR process.
- *Standard users.* Standard users are the workers who carry out tasks such as scanning or OCR correction. This role is played by users with some knowledge in the document field but without any responsibility on the management of the system.

Figure 1 shows the overall system architecture. When we defined it, we followed the recommendations of the Workflow Reference Model (van der Aalst and van Hee, 2002), a commonly accepted framework for the design and development of workflow management systems intended to accommodate the variety of implementation techniques and operational environments that characterize this technology. Thus, although we used this architecture for the implementation of a specific system, it can be used in other environments and situations.

As we can see in Figure 1, the identification and authorizing module is in charge of the authentication of the workers who want to use the system. Each user has a system role depending on the tasks he/she is going to work on. In terms of this system role, the authorizing module only provides the user with access to the needed features. Furthermore, the system architecture is composed of a module for each activity carried out during the repository creation.

- *Metadata storage.* This subsystem is in charge of the introduction and storage of the metadata for each document (title, author, year, source, etc.). This task is performed by the advanced users of the system, therefore only they have access to this module.
- *Scanning.* This system provides access to the scanning hardware and software, and it is the responsible for managing the specification of the scanning parameters for each document (for example, options like scanning two pages at the same time, landscape orientation, resolution, number of colours, etc.).
- *OCR.* It provides access to the OCR software that allows the users to obtain the text of the documents and automatically stores it.
- *Correction.* This module provides the reviewer with both the image and the extracted text to carry out the correction to make the necessary modifications.
- *Markup.* It provides the tools used for marking the text with metadata such as the title, author, page, etc.
- *Abstraction.* Given that the system must be generic, it must support indexing several kinds of documents. These documents will be different not only because they may be stored using different file formats (e.g. plain text, XML, etc.), but also because their contents schema may be different (e.g. the author could be an optional attribute in the different schemas). To solve this problem, we have defined an *abstraction* that represents a *document* as a set of *fields*, each one obtained from the text marked in the previous task.
- *Geo-references.* It provides the tools used to detect the geographic references included

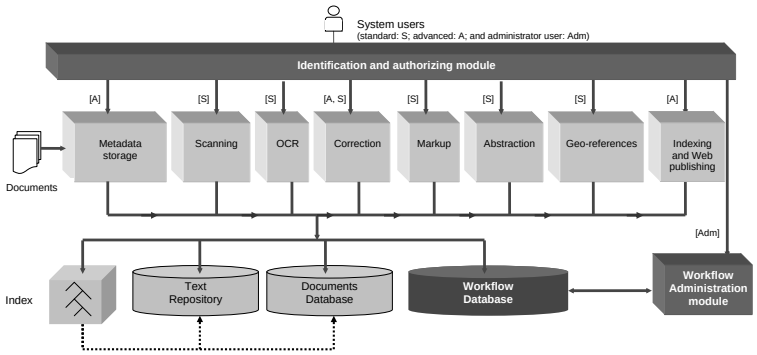

Figure 1: System architecture

into the text of the documents and translate them to a model of the real world (e.g. latitude/longitude coordinates, type of geographic references, etc.). Several proposals have appeared recently to automate this task. However, human performance is not achieved by these proposals. Therefore, both man-made and automated geo-references are possible in the system.

- *Indexing and Web publishing.* Once the document is accepted, this module is in charge of indexing its content using geographic information retrieval techniques.
- *Workflow administration module.* This subsystem is in charge of managing the workflow between all these activities involved in the digitalization. It also provides reporting tools for monitoring purposes.

The system architecture assumes the use of different repositories and databases. The document database and the text repository store the documents and texts extracted from them. An index is built over the document database and text repository to support the search for information. This index, which is described in the next section, combines a textual index, a spatial index, and an ontology-based structure. Finally, the workflow database stores the information about the digitalization chain, with the list of tasks, the state of each document, etc.

4 INDEX STRUCTURE AND SUPPORTED QUERY TYPES

In this section we briefly describe the index structure presented in (Luaces et al., 2008) and the query types that can be solved with it. Figure 2 shows the index structure. The base of this structure is a spatial ontology. This ontology models both the vocabulary and the spatial structure of places for purposes of information retrieval. The structure of the ontology is fixed and therefore our index structure must be constructed ad-hoc for the concrete domain where it will be used.

The main component of the index structure is a tree composed by nodes that represent place names. These nodes are connected by means of inclusion relationships (for instance, Galicia is included in Spain). In each node we store: (i) the keyword (a place name), (ii) the geographic references associated to the place name, (iii) the bounding box of the geometry representing this place, (iv) a list with the document identifiers of the documents that include geographic references to this place, and (v) a list of children nodes that are geographically within this node. Furthermore, an R-Tree is used in each node to improve the performance of the spatial queries.

Two auxiliary structures are used in the index. First, a *place name hash table* stores for each place name its position in the index structure. This provides direct access to a single node by means of a keyword that is returned by a *gazetteer service* if the word processed is a place name. The second auxiliary structure is a traditional inverted index with all the words in the documents that is used to solve textual queries.

Keeping separate indexes for text and geographical scopes has many advantages. First, all textual queries can be efficiently processed by the inverted index, and all spatial queries can be efficiently processed by the index structure. Queries combining textual and spatial aspects are supported, as well. Moreover, updates in each index are handled independently, which makes the addition and removal of data easier. Finally, specific

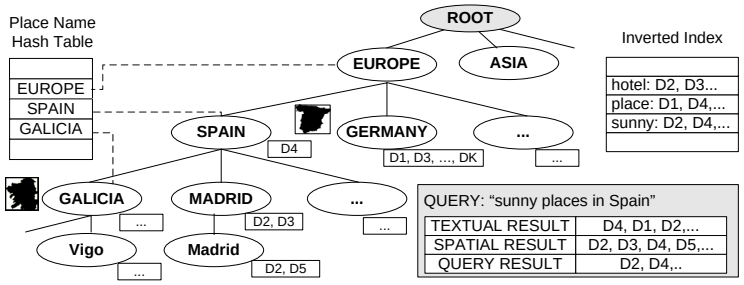

Figure 2: Index structure

optimizations can be applied to each individual indexing structure. On the contrary, the main drawbacks of this structure are: (i) the tree that supports the structure is possibly unbalanced penalizing the efficiency of the system, and (ii) ontologies have a fixed structure and thus our structure is static and it must be constructed *ad-hoc*.

Finally, the most important characteristic of an index structure is the type of queries that can be solved with it. Our index structure support three types of queries: pure textual queries, pure spatial queries, and queries with a textual and a spatial component. In this last type, the spatial component can be given both as a location name and as a geographical area.

Pure textual queries such as "*retrieve all documents where the words* hotel *and* sea *appear*" can be solved by our system because a textual index is part of the index structure. Similarly, pure spatial queries such "*retrieve all documents that refer to the following geographic area*" can also be solved because the index structure is built like a spatial index. Each node in the tree is associated with the bounding box of the geographic objects in its subtree. Hence, the same algorithm that is used with spatial indexes can be used with our structure.

Furthermore, the index structure that we propose can be used to solve queries that involve a textual and a spatial component. In this case, the textual index is used to retrieve the list of documents that contain the words, and the spatial index structure is used to compute the list of documents that reference the geographic area. The result to the query is computed as the intersection of both lists. In the case of queries such as "*sunny places in Spain*" (see Figure 2), our system uses a gazetteer service to discover that *Spain* is a geographic reference and then it uses the *place name hash table* to retrieve the index node that represents *Spain*. Thus, we save some time by avoiding a tree traversal.

Another improvement over text and spatial indexes is that our index structure can easily perform query expansion on geographic references because the index structure is built from an ontology of the geographic space. Consider the following query "*retrieve all documents that refer to Spain*". The query evaluation service will discover that Spain is a geographic reference and the place name index will be used to quickly locate the internal node that represents the geographic object *Spain*. Then all the documents associated to this node are part of the result to the query. Moreover, all the children of this node are geographic objects that are contained within Spain (for instance, the city of Madrid). Therefore, all the documents referenced by the subtree are also part of the result of the query. The consequence is that the index structure has been used to expand the query because the result contains not only those documents that include the term *Spain*, but also all the documents that contain the name of a geographic object included in Spain (e.g., all the cities and regions of Spain).

5 CONCLUSIONS AND FUTURE WORK

The creation of a document repository is not a simple process. It requires the coordination of people and tools to carry out every activity that is part of the process. This process is even more complicated when the geographic references included in the text of the documents must be obtained and translated to a model of the real world. For all this process to be correctly and efficiently made, it is necessary the use of support tools that facilitate the work of each participant and ensure the quality of the obtained results.

The proposed workflow strategies and system architecture support the control and coordination

of people and tasks involved in the digitalization process. The use of this architecture automates the completion of activities that are prone to error and optimizes the performance of the process and the quality of the obtained results. This architecture was defined following the recommendations of the Workflow Reference Model. This system was built as a web application that provides an integrated environment for the execution of all the tasks.

Furthermore, the index structure integrated in the document management system combines a textual index, a spatial index and an ontology-based structure. Finally, new types of queries can be solved with this index structure.

We are currently finishing a prototype of the system. After that, we plan on using it in a real scenario and evaluate its performance. Future improvements of the workflow process and the index structure are possible. First, we plan to include other types of spatial relationships in the index structure in addition to inclusion (e.g. adjacency). These relationships can be easily represented in the ontology-based structure an the index structure can be extended to support them. Another line of future work involves exploring the use of *Toponym Resolution* techniques to improve the task in charge of obtaining geo-references. Finally, it is necessary to define algorithms to rank the documents retrieved by the system. For this task, we must define a measure of spatial relevance and combine it with the relevance computed using the inverted index.

REFERENCES

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Chen, Y.-Y., Suel, T., and Markowetz, A. (2006). Efficient query processing in geographic web search engines. In *SIGMOD Conference*, pages 277–288.

Fischer, L. (2003). *Workflow handbook 2003*. Future Strategies Inc., USA.

Gaede, V. and Günther, O. (1998). Multidimensional access methods. *ACM Comput. Surv.*, 30(2):170–231.

Global Spatial Data Infrastructure Association (2008). Online documentation. Retrieved March 2008 from http://www.gsdi.org/.

Hollingsworth, D. (1995). Workflow management coalition - the workflow reference model. Technical report, Workflow Management Coalition.

ISO/IEC (2002). Geographic Information – Reference Model. International Standard 19101, ISO/IEC.

Jones, C. B., Abdelmoty, A. I., Fu, G., and Vaid, S. (2004). The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the 3rd Int. Conf. on Geogr. Inform. Science*, volume 3234 of *LNCS*, pages 125 – 139.

Lieberman, M. D., Samet, H., Sankaranarayanan, J., and Sperling, J. (2007). STEWARD: Architecture of a Spatio-Textual Search Engine. In *Proceedings of the 15th ACM Int. Symp. on Advances in GIS (ACMGIS07)*, pages 186 – 193. ACM Press.

Luaces, M. R., Paramá, J. R., Pedreira, O., and Seco, D. (2008). An ontology-based index to retrieve documents with geographic information. In Ludaescher, B. and Mamoulis, N., editors, *Proc. of the 20th International Conference on Statistical Scientific Database Management (SSDBM'08) - LNCS*, volume 5069, pages 384–400, Hong Kong, China.

Martins, B., Silva, M. J., and Andrade, L. (2005). Indexing and ranking in Geo-IR systems. In *GIR '05: Proceedings of the 2005 workshop on Geogr. Inform. retrieval*, pages 31–34, New York,USA. ACM Press.

Open GIS Consortium, Inc. (2003). OpenGIS Reference Model. OpenGIS Project Document 03-040, Open GIS Consortium, Inc.

Places, A. S., Brisaboa, N. R., Paramá, J. R., Pedreira, O., and Seco, D. (2007). Managing the workflow of massive feeding of digital libraries. *Research in Computer Science*, 32:352–362.

Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. (2005). Spatio-Textual Indexing for Geographical Search on the Web. In *Proceedings of the 9th Int. Symp. on Spatial and Temporal Databases (SSTD)*, volume 3633 of *LNCS*, pages 218 – 235.

van der Aalst, W. and van Hee, K. (2002). *Workflow management: Models, methods, and systems*.

Worboys, M. F. (2004). *GIS: A Computing Perspective*. CRC. ISBN: 0415283752.