Annotating Digital Documents for Improved Search Results in a Document Repository

Michael K. Baldwin
Department of Computer Science

Abstract

Since annotation is already one of the most commonly occurring activities involving digital documents [1] there is potentially a large amount of extra, user added, information that could be used to improve or provide more powerful and accurate searching methods. This work seeks to develop the technologies, including an annotation tool and a specialized document repository, that are needed to demonstrate this capability. The annotation tool will allow a user to produce XML standoff markup, i.e. markup stored separately from the original document, [2] which will then be stored in the repository along with the document. By utilizing a distributed document annotation process many users can annotate documents for submission to a single central repository that can then use the user submitted metadata to provide more powerful and flexible searching. Search queries can be specified using traditional keyword based approaches or by more powerful query language such as XPath.

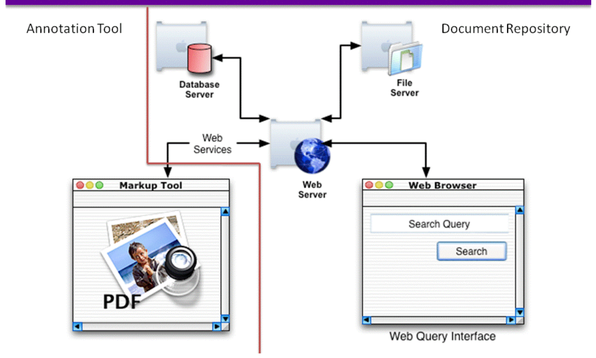
Introduction

Traditional document repositories primarily utilize simple keyword based search techniques, with keywords being entered by people who may not have read the document. Since people typically annotate documents by hand as they are reading them more powerful and meaningful searching could be possible if these annotations could be captured in a structured digital format [1].

This work seeks to address this by providing a tool for producing structured annotations that would be useful for searching. A document repository is also being developed to take these annotations and allow a user to search for documents using this user submitted information.

By using information submitted by actual users we believe that improved search results can be obtained because of the addition of information from users who have read the documents and can identify the most important topics.

System Architecture



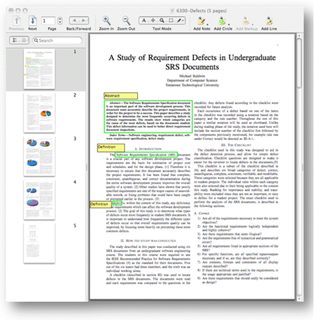
System Architecture

The system consists of two distinct parts: the annotation tool and the document repository. The annotation tool is a desktop tool written in Objective-C for the Mac OS X platform, and is based on an existing tool. The document repository is a completely new system and is designed to be a web application. The system is being implemented using the Apache web server, MySQL 5.1 database server, and PHP. The web server will communicate with the database server to store and retrieve annotation information as XML. Documents will be stored on a file server which can either be on the same system as the web server or on a different server with which the web server can communicate. A user searching the repository can access the system from any platform using any standards compliant web browser.

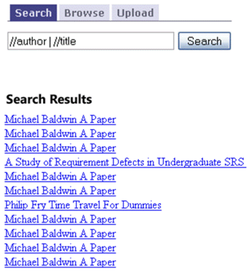
Annotation Tool

An annotation tool based on an existing open source application, called Skim, allows user to add annotations to a PDF document and then export the annotations as XML.

Annotations are added by directly drawing boxes or other shapes over the text of the PDF document. An annotation can then be associated with one of the previously drawn shapes.



Document Repository

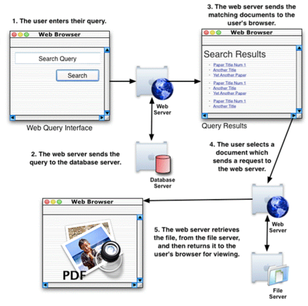


The document repository is a web-based application which provides an interface for users to upload documents and annotation files, browse existing documents, and search for documents using XPath. The search feature will allow users familiar with the annotation format to enter their own custom XPath expressions and provide predefined search terms.

Search Process

When searching for documents in the repository the user will go through the following steps:

1. The user enters a query into the search box.
2. The application will query the database for matching documents.
3. All matching documents will be displayed.
4. The user will select the document they would like to view.
5. The selected document is displayed in the user's browser.



Uploading Documents

Once a document has been annotated and the markup file has been exported the user can then upload the document and annotation file into the repository. The user will also be asked for the document title, author, and an external URL where the document can be found. The PDF document will be saved to the file server and the XML and other information will be stored into the database.

Current Progress

Currently the annotation tool is functional and can export the required XML metadata file for the document repository, but the XML for annotations must be manually created by the user in a text note. The document repository allows for uploading and browsing documents that are in the database and for searching for documents by entering XPath expressions manually. The current search functionality requires detailed knowledge of the XML format and XPath and could be vulnerable to malicious queries since users can enter any valid XPath expression.

Future Work

There is much work remaining on this project to complete both the document repository and annotation tool. The annotation tool needs to be modified to allow annotations to be added by selecting an annotation type and placing a box around the text that the annotation applies to. Ideally the annotation tool will be able to communicate directly with the repository, via web services, to automatically upload and retrieve annotations. The search feature in the repository needs predefined search queries to make the system easier to use for novices and improved security for XPath expressions entered by advanced users. Another area of work for the future could also be the development of a new custom cross-platform annotation tool.

Conclusion

Most current document repositories rely on keyword based approaches for their search functions. These keywords are typically entered based on information provided by authors and not others who have actually read the document. By allowing persons who have read the document provide annotations marking the most important parts of the document it should be possible to improve the quality of the search results users receive.

References

[1] A. J. B. Brush, D. Bargeron, A. Gupta, and J. J. Cadiz, "Robust annotation positioning in digital documents," in CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems. New York, NY, USA: ACM, 2001, pp. 285–292.

[2] Thomas, P. L., and Brailsford, D. F. Enhancing composite digital documents using xml-based standoff markup. In DocEng '05: Proceedings of the 2005 ACM symposium on Document Engineering (New York, NY, USA, 2005), ACM Press, pp. 177_186.