Intelligent Search in eBook Repository

Waris Ali and Ivan Koychev[1]

Faculty of Mathematics and Informatics, University of Sofia, Bulgaria
a.waris@gmail.com, koychev@fmi.uni-sofia.bg

Abstract. Nowadays a lot of systems are available for searching e-books, but the ways to retrieve relevant and meaningful information are difficult. To cope with this problem of retrieving relevant information efficiently we proposed a system, which allows user to search e-books using semantically enriched queries. Our approach not only facilitates users for intelligent e-book search but it also facilitates user in locating their desired information inside the e-book.

Keywords: electronic books, intelligent search, ontology, query formulation

1 Introduction

Advent of the internet and computer technologies has revolutionized almost every possible aspect of our life. Most of the online available material is in the form of text and any one can publish his/her text in order to make it available publically. So aca-demicians take advantage of this and started to publish their articles and books at al-most zero cost. First effort towards development of eBook was started in 1971 by Michael Hart at University of Illinois under the project "Gutenberg" [1] and mission of this project is to encourage creation and distribution of e-books.

According to the Association of American Publishers, an eBook is "a Literary Work in the form of a Digital Object consisting of one or more standard Unique Iden-tifiers, Metadata, and a Monographic body of content, intended to be published and accessed electronically"[2]. The main idea behind representing a book electronically is not to replace paper books with e-books but to enhance the reading activity. The future of eBook is more than the digital form of a paper book - its going to include multimedia within the text. The SmartBook project [5] further enhance this idea aiming to develop an intelligent, community-centred framework for authoring and experiencing of a new generation of 'smart' books - e-books that are evolving, highly interactive, customisable, adaptable, semantically rich, and furnished with a rich set of collaborative authoring and reading support services [5] [6].

In the last few years, the electronic resources are growing rapidly. And the pace of this growth is more than the Moore's Law, so it is challenging and time consuming to retrieve the relevant information according to users' information needs. Majority of current information search systems and book searching systems are based on using keyword search techniques does not have meaning and semantics [3]. Usually user query do not specify meaning or context according

[1] Also associated with Institute of Mathematics and Informatics - BAS.

to user's information require-ments, so it is the goal of the searching systems is to search for a set of documents that will provide him/her necessary relevant information.

Main focus of our research is efficient and intelligent search in e-books repository. To make our search more intelligent and context aware we had to make use of semantic knowledge stored in ontologies. Moreover our research's focus is not only searching across the collections of documents but it also comprises of efficient semantic aware search inside the book or other digital documents.

2 Electronic Book Search Architecture

Our proposed electronic book search architecture is shown in Figure 1, which is based on Lucene search engine. Lucene provides a library for document indexing and searching that can be integrated and customizable with other applications [9] [10]. There are many applications that are using Lucene for searching and indexing pur-poses but here we mention few names: DSpace, MIT's Open-Courseware, SnipSnap, Eclipse IDE, Encyclopedia Britannica, Eyebrowse and Epiphany Web Browser etc. Our proposed architecture is composed of two main modules, query formulation and search engine. Here we make an assumption that documents/ebooks are already stored in local repository and also crawled from distributed repositories. Details of both modules of our proposed architecture are given below.

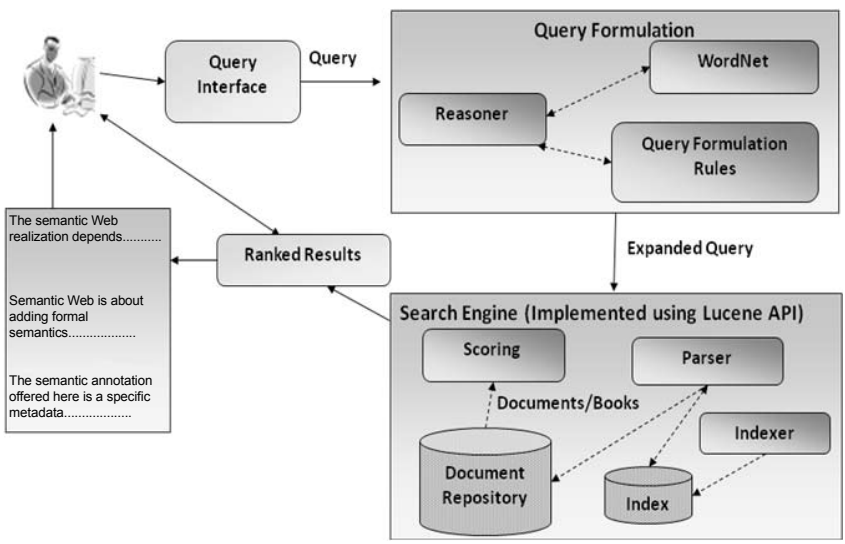

Fig. 1. Proposed eBook search architecture.

2.1 Query Formulation

First step towards the achievement of our research goal is to expand user's query with the addition of semantically related concepts. Reasoner is the core component of query formulation module which is responsible to find concepts semantically related to the queried concept from the WordNet ontology by

using certain predefined formulation rules on the basis of certain relationships like synonym, hypernym and hyponym etc. [7] [8]. Semantically enriched expanded query is formulated by adding newly reasoned concepts along with the user's provided keywords/phrases. Finally this expanded query is submitted to search engine for searching electronic documents and books. In this way we can say that query formulation module is pre processing step before sending query to search engine. Query formulation process gives user more chances to search more relevant documents to his/her information needs.

Ontology: In our research, ontology is used to exploit the relationships between the concepts to find the related concepts. An initial definition was given by Tom Gruber: "Ontology is an explicit specification of a conceptualization". In computer science we can say that ontology is a model to represent domain concepts and relationships between them, and used for reasoning about the concepts of that domain.

2.2 Search Engine

Search engine module is the heart of our proposed architecture and its implemented using Lucene search library which is a fast and flexible java library and can be easily integrated. Main components of search engine module are indexer, query parser and ranker. Document and index repositories are also part of this module.Whenever a new documented is added to the document repository, indexer creates its index and stores it in the index repository along with mapping to the document.

Once semantically enriched query is forwarded to search engine by the query formulation module, first of all query parsing is done and then these parsed query con-cepts and phrases are matched with the already created document index and on the basis of this matching document are selected from document repository. Once the documents are selected against the user's query then it is the responsibility of Scoring component to compute the relevance score for each of the retrieved document and ranked documents are displayed to user as an output.

Efficient Semantic Aware Search inside Ebooks: As we know that even if we have a set of documents related to our information needs, it is very tedious and time con-suming to locate where the relevant information lies in those particular documents. Hence this issue is of great importance to solve, so in recent years some efforts have been put towards this issue like passage retrieval [11] content-based document re-trieval [12]. In our proposed approach we use semantic annotations of documents for the locating information within documents efficiently that facilitate user in finding where the desired information lies inside the documents.

3 Conclusion

User needs more efficient and intelligent ways to access the digital/ebook repositories for fulfilling his/her information requirements. So as a first step towards this we use WordNet ontology for construction of semantically enriched queries to satisfy user in-formation needs intelligently. WordNet is a general ontology; hence there is need for the development of domain specific ontologies. Our proposed

approach is first step towards the development of proper intelligent search system that is able to retrieve and rank documents by their semantic properties. Our system allows user to search the ebooks with limited information. Another important feature of our system is that it will also provide in-book context aware search. In future to make our search more efficient and context aware we will take advantage from the annotations/tags and comments given by the user while he/she is viewing those documents/ebooks. As the tags are assigned by the users of that particular domain so those tags will reflect the context more precisely. We are also wanted to keep user's interests and searching behaviour that would be helpful in finding the context of user's queries and in recommending related documents to his/her search interests.

Acknowledgements. This research is supported by the SmartBook project, subsidized by the Bulgarian Na-tional Science Fund, under Grant D002-111 /15.12.2008

References

1. Gutenberg: The History and Philosophy of Project Gutenberg by Michael Hart, http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart3
2. Bolick, R.: Publishers' Requirements for Digital Rights Management. In: W3C Workshopon Digital Rights Management for the Web, France (2001)
3. Lanin, V., Lyadova, L.: Intelligent Search and Automatic Document Classification and Cataloging Based on Ontology Approach. International Journal "Information Theories & Applications", 14, 25-29 (2007)
4. Roo, S.S.: Electronic book technology: an overview of the present situation, Library Review Journal. 53, 363-371 (2004)
5. Koychev, I., Nikolov, R., Dicheva,D.: SmartBook – a vision for the future e-book, Advances in Bulgarian Science, 3 (2009)
6. Koychev I., Dicheva D. and Nikolov, R.: SmartBook: Semantics Inside – to appear in Serdica Journal of Computing.
7. Ali, W., Khan, S.: Ontology Driven Query Expansion in Data Integration. In: 4th International Conference on Semantic Knowledge and Grid, pp. 57-63. IEEE Press, China (2008)
8. Grootjen, F.A.,van der Weide, Th. P.: Conceptual Query Expansion. Journal of Data Knowledge and Engineering. 56, pp. 174-193 (2006)
9. Apache Lucene - Overview, http://lucene.apache.org/
10. Parsad, A.: Lucene Search Engine: An Overview. In: DRTC-HP International Workshop on Building Digital Libraries using DSpace, India (2005)
11. Hareast, M.A.: TileBars: Visualization of Term Distribution Information in Full Text Information Access. In: ACM SIGCHI Conference on Human Factors in computing Systems, pp. 65-71 (1995)
12. Harper, D.J., Koychev, I., Sun, Y. and Pirie I. (2004). Within-document Retrieval: A User-Centred Evaluation of Relevance Profiling, Journal of Information Retrieval, 7, 265-290, 2004, Kluwer Academic Publishers.