INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH & DEVELOPMENT

ISSN 2278 – 0211 (Online)

Focused Retrieval of E-books using Text Learning and Semantic Search

Kirti D. Pakhale
PG Student, DYPCOE, Akurdi, Department of Computer Engineering, University of Pune, India
S. S. Pawar
Assistant Professor, DYPCOE, Akurdi, Department of Computer Engineering, University of Pune, Maharashtra, India

Abstract:
There are many online digital libraries (DLs) containing books, authors and subjects' data, which are accessed via internal search services as well as external web sites such as Google, Yahoo. Due to a continuous rising of heterogeneous data on the Web it is difficult for users to access relevant information online. Digital Libraries face similar challenges due to fast growth of available electronic data. Increase availability of users to access full-text of digitized books on the Web it prompts Digital Libraries to enhance usability and the searching techniques to obtain highly relevant and more focused results (e.g. Books or answers to a user query). The traditional Information Retrieval methodologies generally retrieved documents by matching terms in documents that contains user specified keywords and the term with higher frequency. The proposed system solves the problem by using ontology, metadata and semantic search. This model consists of identifying concepts in users' query and expands them by creating ontology from text learning. The proposed system aims to improve usability, relevancy ratio with respects to accessing books from Digital Libraries based on user query. The proposed system will retrieve particular chapters based on the user query.

Keywords: *Information Retrieval, Ontology, Semantic Search, Text Learning*

1. Introduction

There are many online DLs that collect, organize, and offer information about books. If a user wants to access this content, they can do it in several ways. One way is to visit the digital library website directly and browse or search using the internal search interface. Second way is through web search engines (WSEs) use for finding book-related content. Now a day's books are most widely used for knowledge, entertainment and information. For the last few years a lot of activities took place in the field of Digital Libraries by using the new information technologies so as to ease access of information stored in DLs. Information Retrieval System (IRS) is a process of searching, representing and retrieving the most relevant documents as per the user query. A digital library is a type of IRS in which collection of books, documents are stored in digital formats and accessed remotely via computer networks by multiple users at any time without considering physical or geographical boundaries [1]. To improve the usability and relevancy within the DLs motivates researchers' to develop new techniques to achieve more focused and relevant result to the user query. The recent research on "Focused Search'' aims to contributes in reducing such cognitive load on the user by locating relevant content from irrelevant content within a document. Focused retrieval techniques allow users to gain direct access to parts of books (of potentially thousands of pages or particular chapters from book) relevant to the information needed.

Traditional information system retrieved the document by matching terms in documents with those of a user query. It depends on the term frequency and term weight analysis of the text document. It retrieves documents that contain keyword specified by user and the term with higher frequency [2]. These approaches have some limitation in extracting semantically similar terms that represent the similar meaning in the documents. Many documents convey desired information on the basis of semantic without containing these keywords. To get better relevancy of the retrieved documents and the usability. In this paper we focus on the efficient information retrieval using ontology as a controlled vocabulary to expand the input string. The proposed system represents ontology-based framework. Ontology is a collection of concepts and their interrelationships. The use of ontologies for information retrieval is discussed in [3]. To efficiently retrieve books relevant to user query, the proposed model in this paper provides semantic search used to extend traditional keyword search with extracted and inferred information using ontology. Semantic representations are used to map the documents with the concepts and the similarity measures are calculated appropriately to retrieve most relevant results.

2. Related Work

In traditional (flat) information retrieval, the results are typically presented as a list of matching documents. In case of books, the user needs to know, the specific location of text in the books relevant to the query, so there should be further information about relevant

sections or chapters. Philipp Dopichaj explained the working of two online library services named as Books24x71 (launched in 1999) and Safari2 (launched in 2001) that offer full text search for their online books [4]. They presented a list of relevant books and the titles of the most relevant sections of that book. In both the services, book results are overlapped to the content. Simone Marinai et.al. presented a full text of book accessed [5], in which they used the concept of Document Image Retrieval (DIR) that allows users to retrieve digitized pages on the basis of layout similarities and to perform textual searches on the documents without relying on Optical Character Recognition (OCR). Also they used different techniques on page layout to retrieve the relevant pages without considering its semantic meaning. Jin Young Kim et.al. introduced Open Library (OL). The OL provides several features for searching and exploring the stored book records explained in [6].

So to improve online services and the e-book search, we propose a system that contains ontology and semantic searching. Most of the current information systems and search applications use ontologies as a database of additional knowledge representation to perform faster content search access and relevant information to the user query.

2.1. Ontology

Ontologies offer an efficient way to reduce the amount of information overload by encoding the structure of a specific domain and offering easier access to the information for the users. Ontologies play an important role in providing a controlled vocabulary of concepts, each with an explicitly defined and machine understandable semantics. Due to increase availability of information on Internet IRS started to be applied to large volumes of data. In [7], it is explained how ontologies were developed in the EU Semantic Web project: SPIRIT. The query expansion techniques presented in this paper were based on domain ontology. Although most concept-based IRS used the WorldNet as a controlled vocabulary to expand queries, in our proposed system we use ontology as a controlled vocabulary for query expansion. In recent years, it was explored that even with the best indexing techniques; a good precision in search results has not been obtained yet. The evolution of Web 3.0: Semantic Web proposed to clarify the meaning of resources by annotating them with metadata i. e. data over the data. By associating metadata to resources, semantic searches can be significantly improved as compared to traditional searches. The main advantage of the Semantic Web is to enhance search mechanisms with the use of ontologies and allow users to use natural language to express what he wants to search. Several proposals of semantic search systems exist now a day. The main idea is expanding queries with the semantics of the words to achieve better recall and precision.

2.2. Ontology Development

To build domain ontologies from the knowledge requires much time and many resources. Ontology learning can be used as the set of methods and techniques for building ontology from scratch, or enriching, or adapting an existing ontology in a semi-automatic fashion using several sources. Ontology learning is a wide domain of research that consists of extending an existing ontology with additional concepts and relations and placing them in the correct position in the ontology, resolving inconsistencies that appear in ontology with the view to acquire consistent (sub) ontology and ontology population is adding new instances of concepts into the ontology. Alexander Maedche and Steffen Staab (2001) distinguish different ontology learning approaches focused on the type of input used for learning. With respect to that they propose the following classification: ontology learning from text, from dictionary, from knowledge base, from semi-structured schemata and from relational schemata [8]. Depending on the different assumptions regarding the provided input data, ontology learning can be addressed via different tasks: learning the ontology concepts, learning the ontology relationships between the existing concepts, learning both the concepts and relations at the same time, populating an existing ontology or structure, dealing with dynamic data streams, simultaneous construction of ontologies giving different views on the same data, etc. Buitelaar et al. (2005) found information on ontology learning from text. Different ontology learning tools was introduced such as KEA (Jones and Paynter, 2002), OntoLearn (Velardi et al., 2005), Welkin (Alfonseca and Rodriguez, 2002), and Text2Onto (Ciniamo and Volker, 2005) [9].

The proposed system is focused on creating ontology for “Computer Science’’ domain. Ontology based semantic search model is used to ease the access of e-books and enhance efficiency and accuracy of information retrieval. Ontology is created using Text2Onto tool. It is a framework for ontology learning from textual resources. Text2Onto can represent the learned knowledge at a meta-level in the form of instantiated modelling primitives within Probabilistic Ontology Model (POM); it remains independent of a concrete target language and able to translate the instantiated primitives into any knowledge representation formalism.

2.3. Ontology Transformation’s

Ontology translation refers to the process of changing the formal representation of the ontology from one language to another. Ontology transformation includes its expression in a different ontology language, or a reformulation in a restricted of a language. It is useful for solving heterogeneity problems. Ontology information translated using various ontology representation languages such as RDF (Resource Description Framework, OWL (Web Ontology Language, F-Logic [10]. The OWL [11] describes classes, properties, and relations among these conceptual objects. In Text2Onto it uses translation-based approach to knowledge engineering and defines the relevant modelling primitives in the MPL (Modelling Primitive Library). Ontology writers are then responsible for translating instantiated modelling primitives into a specific target knowledge representation language. The modelling primitives use in Text2Onto like concepts (CLASS), concept inheritance (SUBCLASS-OF), concept instantiation (INSTANCE-OF), properties/relations (RELATION) domain and range restrictions (DOMAIN/RANGE), mereological relation, equivalence [12].

3. Implementation Details

The goal of the proposed system is to design ontology from text learning and enhancing the results using lucene indexing and semantic search. Text2Onto [13], which produce the OWL files. It is open source and targets data driven change discovery using an incremental ontology learning strategy from text. In proposed model first it calls Text2Onto providing a set of input documents (Books) in PDF, HTML or plain text, then Text2Onto applies its concept extraction algorithm and calculates the relevance values for each of the found concepts. It exports the concepts into an OWL file (an ontology file format). The system will read this OWL file and presents the extracted concepts.

3.1. System Architecture

The Fig.1 shows the basic architecture of the system. Proposed system uses different algorithms for generating ontology from text.

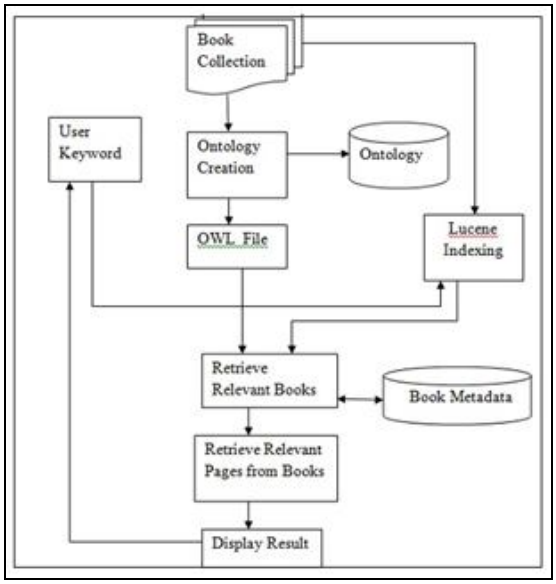

Figure 1: System Architecture

3.2 Algorithms

Different modelling primitives are used by Text2Onto tool to develop ontology. To calculate these modelling primitives different algorithms are used following section describe the different algorithms [12].

3.2.1. Concepts

For extracting concepts following algorithms are used RTFConceptExtraction and TFIDFConceptExtraction It uses different measures like Relative Term Frequency (RTF), TFIDF (Term Frequency Inverted Document Frequency), For each term, the values of these measures are normalized into the interval [0..1] and used as corresponding probability in the POM.

3.2.2. RTF Concept Extraction

It calculates Relative Term Frequency which is obtained by dividing the absolute term frequency (number of times a term t appears in the document d) of the term t in the document d divided by the maximum absolute term frequency (the number of times any term appears the maximum number of times in the document d) of the document d.

$$tf(t,D) = \frac{\text{absolute term frequency}}{\text{Maximum absolute term frequency}}$$

3.2.3. TFIDF Concept Extraction

Equation (1) calculates term frequency inverse document frequency which is the product of TF (term frequency) and IDF (Inverse Document Frequency). Equation (2) obtained IDF by dividing the total number of documents by the number of documents containing the term, and then taking the log of that quotient.

$$tf\_idf=(t,d,D) * idf\ (t,d) \quad (1)$$

Where

$$idf\ (t,D)=\log |D|/df(t) \quad (2)$$

$|D|$= No. of all documents.

idf(t) = No. of all documents containing term.

3.2.4. Subclass-of Relations

Subclass-of relations identification involves several algorithms which use hypernym structure of WordNet. We uses WordNet2.0 for finding subclass_of relation. These algorithms depend on the result of concept extraction algorithms. Relevance calculated using WordNetClassificationExtraction. It extracts subclass-of relations among the extracted concepts identifying the hpernym structure of the concepts in WordNet. Relevance is calculated as if a is a subclass of b, then

$$\text{Relevance} = \frac{\text{No. of synonyms of a for which b is a hypernym}}{\text{No. of synonyms of a}}$$

3.2.5. Instance-of Relations

Lexical patterns and context similarity are taken into account for instance classification. A pattern-matching algorithm similar to the one use for discovering mereological relations is also used for instance-of relation extraction.

Equivalence and equality: The algorithm calculates the similarity between terms on the basis of contextual features extracted from the corpus.

3.3. Semantic Search

In proposed model we used WordNet 2.0 as the taxonomy to calculate semantic similarity between words. WordNet is an online English lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. Various kinds of relationship exist in the WordNet taxonomy, which can be categorized as semantic and lexical relationship. WordNet allows extending the searched terms with three main types of relationships: synonyms, hypernyms and hyponyms. Searching for related terms increases the chances of finding a matching within the index.

4. Results and Discussion

4.1. Results and Comparative Analysis

The Fig.2 shows the retrieval accuracy between Lucene and Lucene + Ontology. Here, keywords are shown on X-axis and accuracy percentage is shown on Y-axis.

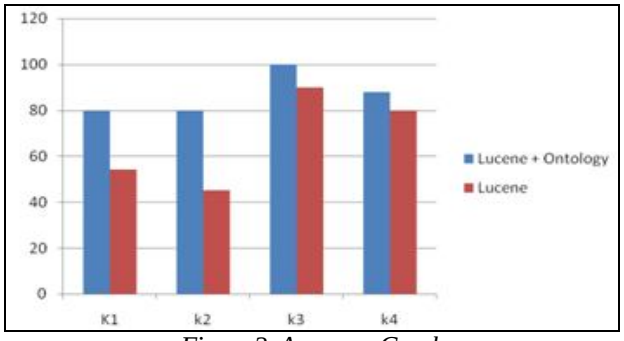

Figure 2: Accuracy Graph

The Fig.3 shows time analysis graph for the same. Here initial time is required for ontology download is high afterword the retrieval time required for Lucene + Ontology is same as Lucene.

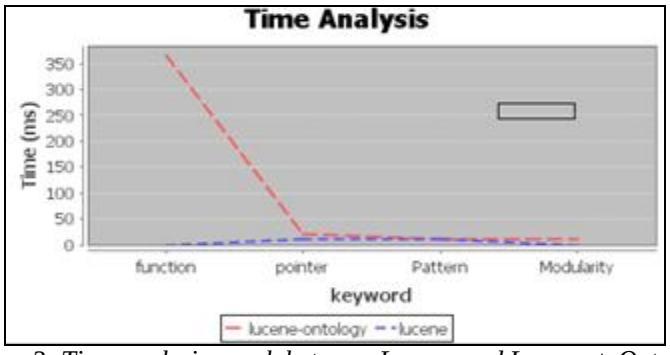

Figure 3: Time analysis graph between Lucene and Lucene + Ontology

5. Conclusion and Future Work

Now days, users are keen to access e-book to obtain knowledge from the digital libraries. The proposed system presents a semantic retrieval framework on the bases of ontology and its application in accessing books from Digital Library for specific domain. It includes all the aspects of semantic retrieval, ontology development, information extraction, semantic search and retrieval. The proposed system is designed to retrieve the more relevant book with respect to user query and also provide easy access of books by locating its chapter number with respective page numbers. Semantic Search is necessity of the word as it is very difficult to get the relevant information from the information ocean. Several algorithms are designed in this domain and researchers are still designing new one. The system can be further refined to improve the effectiveness of IR systems by test it on larger scale and by enhancing the semantic search techniques.

6. Acknowledgment

The authors would like to thank the publishers, researchers for making their resources available and the teachers for their valuable guidance and to the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

7. References

1. Greenstein, Daniel I., and Suzanne E. Thorin. The digital library: A biography. Washington, DC: Digital Library Federation, Council on Library and Information Resources, 2002.
2. Salton, Gerard, and Christopher Buckley. "Term-weighting approaches in automatic text retrieval." Information processing & management 24.5 (1988): 513-523.
3. Minghong Liao, Andreas Abecker, Ansgar Bernardi, Knut Hinkelmann, and Michael Sintek, "Ontologies for knowledge retrieval in organizational memories." Proceedings of the Learning Software Organizations (LSO'99) workshop, Kaiserslauten, Germany. 1999.
4. Dopichaj, Philipp."Element retrieval in digital libraries: Reality check." Proceedings of the SIGIR 2006 Workshop on XML Element Retrieval Methodology. 2006.
5. Marinai, Simone, Emanuele Marino, and Giovanni Soda. "Exploring digital libraries with document image retrieval." Research and Advanced Technology for Digital Libraries. Springer Berlin Heidelberg, 2007. 368-379.
6. Kim, Jin Young, Henry Feild, and Marc Cartright. "Understanding book search behavior on the web." Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012.
7. Fu, Gaihua, Christopher B. Jones, and Alia I. Abdelmoty. "Ontology-based spatial query expansion in information retrieval." On the move to meaningful internet systems 2005: CoopIS, DOA, and ODBASE. Springer Berlin Heidelberg, 2005. 1466-1482.
8. Philipp Cimiano, Johanna VÄolker, Text2Onto, A Framework for Ontology Learning and Data-driven Change Discovery.
9. Seremeti, Lambrini, and Achilles Kameas. "Tools for Ontology Engineering and Management." Theory and Applications of Ontology: Computer Applications. Springer Netherlands, 2010. 131-154.
10. Kifer, Michael, Georg Lausen, and James Wu. "Logical foundations of object-oriented and frame-based languages." Journal of the ACM (JACM) 42.4 (1995): 741-843.
11. http://www.w3.org.
12. Sonam Mittal, Nupur Mittal, Tools for Ontology Building from Texts: Analysis and Improvement of the Results of Text2Onto, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 11, PP 101-117, May - Jun. 2013.
13. Cimiano, Philipp, et al. "Ontology learning." Handbook on ontologies. Springer Berlin Heidelberg, 2009. 245-267