A Comparison Study Between Generic and Metadata Search Engines in an E-learning Environment

Leyla Zhuhadar, Olfa Nasraoui, and Robert Wyatt
Dept of Computer Engineering and Computer Science, University of Louisville, Louisville, KY, USA
Office of Distance Learning, Western Kentucky University, Bowling Green, KY, USA

Abstract—The huge explosion of the amount of information on the web makes it difficult for online students to find specific information with a specific media format unless a prior analysis has been made. In this paper, we present a comparison study between Generic and Metadata Search engines in a multimedia repository of lectures. Both Search engines were modified to index text, powerpoint, audio, video, podcast, and vodcast lectures. These lectures are stored in a prototype E-learning web-based platform. Each lecture in this platform has been tagged with metadata using the domain-knowledge of these resources. This comparison study was evaluated based on precision.

Index Terms—Information Retrieval, Metadata, Search Engine, Ranking Algorithm, E-learning.

I. INTRODUCTION

Nowadays, the Web contains billions of dynamic resources. Searching for information, which could be a document, an image, an audio, or a video file, a podcast, a vodcast, a blog, etc. from online repositories, such as digital libraries can be a difficult task in IR. Therefore, new techniques in IR emerged to increase the quality of searching. These techniques focused on three areas of interest: the capability of handling a new indexing format (which allows the indexing of images, audio, video, rss, etc.), enhancing the indexing to have a faster response time for queries and the quality of the information retrieval systems.

Only minimal work has been done on proactively merging efforts from Information Retrieval (IR) with Education, such as [23], [5] which is not directly related to the E-learning domain research. The need for this synergy arises from the fact that the number of online students has been growing significantly; nearly 20% of all U.S. higher education students were taking at least one online course in the fall of 2006, and almost 3.5 million students were taking at least one online course during the fall 2006 term; a nearly 10% increase over the number reported the previous year, and an increase of 9.7% growth rate for online enrollments far exceeds the 1.5% growth of the overall higher education student population. These facts are based on a survey that represents the fifth annual report on the state of online learning in U.S. higher education.\(^1\) In this paper, we present a metadata topic-driven search engine that uses Information Retrieval methods to enhance the indexing and searching of learning objects in an E-learning repository. The Search engine is embedded on top of a web-based platform to allow indexing of text, powerpoint, audio, video, podcast, and vodcast lectures. These lectures are stored in a prototype “HyperManyMedia” E-learning web-based platform.

II. PREVIOUS WORK

Recently, two notions of search engines started to gain popularity: the metadata search engine and the focused/topical search engine. The metadata search engine is based on a metadata structure, which is “machine understandable information about web resources”\(^2\). The metadata search engines have been studied intensively in [11], [18], [4], [6], [10], [12], [20]. On the other hand, focused/topical search engines were introduced for the first time in [3] and a great deal of research related to them has been presented in [4], [7], [13], [15], [16], [20], [22]. Our approach uses hybrid “metadata” and “topic-driven” mechanisms and it is capable of indexing and fetching many different media format resources from text, powerpoint, audio, video, podcast, and vodcast. Since our search engine is built on top of the domain-knowledge of (E-learning), this knowledge representation was extracted from the subject area, which combines topics and media formats about online resources. Finally, we mapped this knowledge into “metadata” to enhance the Educational (E-learning) search mechanism.

We have organized the paper as follows. Sections 3 presents our methodology and implementation of a metadata search engine. Section 4 presents our results and evaluation measures. Section 5 includes our conclusions and future works and finally, section 6 provides the references.

\(^1\)http://www.sloan-c.org/publications/survey/survey07.asp
\(^2\)http://www.w3.org/DesignIssues/Metadata
III. METHODOLOGY AND IMPLEMENTATION

A. System Architecture

1) Topic-knowledge Extraction: Given a learning object (lecture) from the “HyperManyMedia” Server, we first locate the learning object (lecture) from the server; then, we extract information about this lecture. For example, what type of format is the learning object? Is it text, power-point, audio, video, podcast or vodcast.

2) Parsing Learning Objects (Lectures) and Adding Metadata: We use this information to embed the metadata (e.g., college name, course name, professor name, media format type). This phase is more relative to our domain-knowledge of these lectures.

3) Re-configuring a Search engine for Multi-media Indexing and Querying: This phase is related to the search engine architecture. Plugins were embedded inside a search engine, which automatically add additional fields (metadata fields). This enhances the search engine’s capability of indexing and searching for metadata.

4) Encapsulating the Metadata Search Engine within the Platform: Finally, we embedded “metadata” search engine in the E-learning platform. The search engine facilitates the searching mechanism for online students. Students could search for a specific college, course name, topic, or media format accurately. Moreover, searching for combinations of results, for example, finding all video lectures related to a specific college is now possible.

B. System Implementation

1) Topic-knowledge Extraction: Our “HyperManyMedia” Server contains more than four hundred lectures from eleven different colleges: “English”, “Social Work”, “History”, “Chemistry”, “Accounting”, “Math”, “Management”, “Consumer and Family Sciences”, “Architect and Manufacturing Sciences”, “Engineering” and “Communication Disorders”. Each lecture is delivered in six different media formats: text, powerpoint, audio, video, podcast, and voodcast. Each lecture is a learning object used in online courses and taught by different professors. Information about each lecture was extracted and saved. This phase was done semi-manual, since the knowledge of each resource in the platform can only be known by web designer, web developer, multimedia editors, moreover, tagging this information inside the learning objects (lectures) is described in the next section.

2) Parsing Learning Objects (Lectures) and Adding Metadata: All webpages (lectures) located on the server were parsed using a java program which parses each webpage to find the specific location for the metadata. The metadata information describes the following criteria: college name, course name, professor name, lecture name, media format type. Fig. 2. shows an example of the added metadata to an English course.

3) Re-configuring Nutch:

- A Brief Overview of Nutch

Nutch\[^3\] is an open-source search engine based on Apache Lucene,\[^4\] which is a scalable Information Retrieval (IR) library that allows indexing and searching capabilities. Nutch has been used in many research applications as a retrieval system in digital libraries [17], [19], [11] and as a web search engine [14], [1]. Nutch searches and indexes components with a powerful fetcher (crawler robot), which is designed to handle crawling, indexing, and searching of several billion frequently updated web pages. It has a modular architecture, which permits developers to design and embed plugins for media parsing, data retrieval, querying, and clustering. Nutch has different types of fields, such as keywords, text, anchor, [1](http://lucene.apache.org/java/docs/).

\[^3\]http://www.nutch.org
\[^4\]Lucene is a high-performance, full-featured text search engine library written entirely in Java. (http://lucene.apache.org/java/docs/)
etc. We added new fields as metadata fields which represent
domain-specific fields, such as, college names, lecture names,
and media types. Nutch can parse many different file formats
such as html, php, doc, pdf, rtf, etc. We wrote our own
plugins that add new types of files to be indexed and
searched, such as file with format of type ppt, mp3, mp4,
xml, php.

- Nutch Implementation
The Nutch search engine was implemented in two stages,
first as a “Generic” search engine; second, as an enhanced
“metadata” search engine.

- Nutch Scoring Algorithm
The Nutch scoring algorithm is inherited from apache lucene,
which is based on a combination of the Vector Space Model
(VSM) and of the Boolean Model. It applies the Boolean
Model first to select the most relevant documents for the
query; then, it uses the Vector Space Model as a content-
based ranking algorithm.

The score of query \( q \) for document \( d \) is related to the
cosine-distance similarity (1) between the document and
query vectors in a Vector Space Model (VSM).

\[
cos(x,x') = \frac{x^T \cdot x'}{\|x\| \cdot \|x'\|} = \frac{x^T \cdot x'}{\sqrt{x^T \cdot x} \cdot \sqrt{x'^T \cdot x'}}
\]

where \( x, x' \) are vector-space representations of
two documents, \( T \) denotes the 'transpose' operator and \( x^T \cdot x' \)
indicates the dot product between two vectors. Nutch uses
several refinements on (VSM) by extending the Boolean
vector model and adding weights associated with terms and
fields as shown in (2). This represents Nutch’s scoring,
which is influenced by the sum of the score for each term
of a query. For each field, the score is the product of the
following factors: Its "tf", "idf", and index-time boost. The
score is computed as following:

\[
score(q,d) = \text{coord}(q,d) \times \text{queryNorm}(q) \times \sum (t f(t \text{ind}) \times \text{idf}(t) \times t . \text{getBoost}() \times \text{norm}(t,d))
\]

- Designing and Embedding the Parser, Indexer, and
QueryFilter Plugins
Plugins were embedded inside Nutch that automatically add
additional fields (metadata fields). This enhances Nutch’s
capability of indexing and searching for metadata-based
retrieval. Four plugins were added (colleges, courses, pro-
fessors, and media-format). For each plugin, three programs
were embedded to boost the parser, indexer, and query filter
in order (e.g., CollegeParser, CollegeIndexer, and College-
QueryFilter).

4) Encapsulating Metadata in the Search engine
within the “HyperManyMedia” Platform: Fig. 3.
presents the “HyperManyMedia” Platform and the encapsu-
lated “Metadata” Search engine. It shows different Metadata
fields (College name, Course name , Professor name, Media
Format).

In Fig. 4, we can see the boosting results of the “Generic”
Search engine which has been calculated for a keyword

“Socialwork”. In this case, only the weight of the content
was considered. Whereas, in Fig. 5, we can see the boosting
results of the “Metadata” Search engine for the same
keyword “Socialwork” which has been calculated based
on the Metadata field (college = Socialwork), which gave
a higher boosting score, as consequence, a higher ranking
score.

IV. RESULTS AND EVALUATION MEASURES
A. Evaluation Methodology

1) Research Question: Will there be an increase
in precision when using the metadata search engine
compared to the generic search engine?

2) Selection of Queries: The first step was selecting
queries. A great deal of research on search engine queries
has found that searchers rarely use Boolean operators [2];
typically, this usage is around 10% [8]. Another study [9]
observed that the highest distribution of the number of terms
in queries range between 1 and 3, and these are primarily
noun phrases. Accordingly, we ran our comparison between
the two search engines (generic) and (metadata) based on
“single-term”, “two-terms”, and “three-term” queries without
Boolean operators.

We selected specific queries from the query logs con-
taining queries submitted to our “HyperManyMedia” search
Fig. 5. “Metadata” Search engine Score for the Query=“Socialwork” engine during two semesters (fall & winter terms 2007-2008). These queries represent the usage of the search engine by online students. First, the query log file was cleaned from irrelevant data and from misspelled terms. Second, only the “Top_list” of most frequent queries was extracted. Finally, these terms were ordered in a descended order.

Subjectivity is one of the elements that should be considered when we evaluate a query. A web resource is related not only to a query but also to a user’s nation, age, gender, career, culture, hobby, etc.[21]. A study of satisfaction factor of users could be considered to support the evaluation of the retrieved results. In our case, the search engine is used primarily by online students who have knowledge of what kind of materials they are looking for (this knowledge was provided by their online faculty), for example, college name, course name, faculty name, lecture title, etc. Therefore, our evaluation methodology relies on the online students’ domain knowledge of the resources they are looking for. Consequently, the subjective factors have a limited influence in our method.

3) Precision : Precision is the ratio of the number of relevant documents to all retrieved documents as shown in (3)

\[
\text{precision} = \frac{\text{number of relevant documents}}{\text{number of retrieved documents}} \times 100 \quad (3)
\]

Where number of retrieved documents is equal to the sum of relevant and irrelevant retrieved documents.

(i) Precision Results:
Of the 100 top_list queries submitted in both search engines (“Generic” and “Meta-data”), we recorded the retrieved and the relevant documents, in addition to the order of their ranking. We repeated the experiment three times based on “single-term”, “two-term”, and “three-term” queries. We considered combined words like “communication disorders” as a single-term since these two words are related.

We conducted a comparison of the two search engines “Generic” and “Metadata” to determine if there were any significant differences in terms of precision.

TABLE I presents the overall precision results for the three experiments. The average precision for each experiment indicates that the “Metadata” search engine outperformed the “Generic” search engine. Given that the reviewers were aware of the “Metadata” mechanism in the search engine, for retrieving the lectures and how the metadata was designed, there is a possibility of bias in this analysis. However, the students who use this platform are also aware of this mechanism. Our goal was to compare and contrast both search engine performances using the most frequently used queries by our online students.

![Overall precision](image)

We conclude that the metadata search engine increases the precision across all numbers of terms used. This in turn answers our research question:

(1) Will there be an increase in precision when using the metadata search engine?

We found that the metadata-driven search engine has a significant impact on the precision with overall precision values equal to 0.810 (for single-term queries), 0.856 (for two-term queries), and 0.925 (for three-term queries), compared to 0.619 (for single-term queries), 0.717 (for two-term queries), and 0.851 (for three-term queries) for the generic search engine, as shown in Fig. 6.

The Generic search engine was able to generate a better precision by a very small percentage \(\approx 3-6\%\) in some queries where we do not have metadata. In these cases, the reviewers selected the metadata which is the closest to the query terms. As a consequence, it skewed the results. However, this encouraged us in two directions; first, to think about adding some new metadata to our platform; and second, to support our idea of the need of creating E-learning ontologies. E.g., the problem of synonymous words could be reduced if we had a semantic search engine. This will be considered in our future work.
V. CONCLUSION AND FUTURE WORK

In this work, we presented a metadata search engine in “HyperManyMedia” E-learning resources. Our results based on precision showed a significant improvement in retrieving relevant resources to the submitted queries when we used the metadata search engine. The evaluation of these results was based on our domain knowledge of these resources. Our Search engine was modified to fit an E-learning repository of multimedia lectures. It is capable of handling new indexing formats (which allows the indexing of powerpoint, audio, video, rss, etc.).

Our future research will focus on designing a hybrid metadata and semantically enriched search engine which will be built on top of the domain-knowledge of (E-learning). The knowledge representation will be extracted from the subject area. Then, we will map these concepts into ontologies using learning objects. The learning objects will contain metadata that will help online students to identify the most suitable resource for a specific lecture. Moreover, these learning objects will be associated to learner’s profiles using personalization. Finally, we will visualize online students communities with their associated learning objects and their relationships.

Acknowledgment
This research was partially supported by National Science Foundation CAREER Award IIS-0133948 to O. Nasraoui.

VI. REFERENCES


