ABSTRACT

AUTOMATED DISCOVERY, CATEGORIZATION AND RETRIEVAL OF PERSONALIZED SEMANTICALLY ENRICHED E-LEARNING RESOURCES

Leyla Zhuhadar

December 4, 2009

In the past 30 years, Information Retrieval (IR) was considered as a field of interest mainly to librarians and information experts. With the invention of the World Wide Web (WWW), IR evolved and new techniques started to develop, such as tools and applications for retrieving data in the context of the Web. Today, the Web can be considered as the aggregate of huge hyperlinked digital libraries with data scattered on billions of dynamic resources. Searching for information, which could be a document, image, audio, video, podcast, vodcast, or blog, etc., from these libraries has become an increasingly difficult task in IR. Therefore, several efforts have emerged to increase the quality of organizing and searching this information. These efforts focus on three areas of interest: (a) the capability of handling a new indexing format (which allows the indexing of images, audio, video, RSS, etc.), (b) enhancing the indexing to have a faster response time for queries, and (c) improving the quality of the information retrieval system.

This dissertation falls within what we consider to be an emerging area in the Colliding Web Sciences that we refer to as IR for E-learning; it merges Web Information Retrieval with E-learning. The need for this synergy arises from the fact that the number of online students has been growing significantly for the last couple of years and the generic Web Information Retrieval methods have either maintained an emphasis on serving the general population, or have been lagging in integrating the power of adaptive semantics and personalization via knowledge discovery into real-life working E-learning applications. In this dissertation, we propose an integrated and working E-learning
search system for retrieving personalized semantically enriched learning resources; our system is based on five components.

The first component of this proposed system is related to Information Filtering, which despite its improvements for the past decade has focused on models that are “good for all users” and not for a specific user. The enormous increase of information on the Web led the information retrieval community to strive toward changing the concept of “good for all” to “good for everyone.” This, in turn, popularized personalized semantic search engines and semantically enhanced recommendation systems. Thus, there was a need for an infrastructure that can provide, manage, and collect data that permit high levels of adaptability and relevance to the learner’s profiles. Within this context, this work proposes an architecture divided into four layers: (1) Semantic Representation (knowledge representation), (2) Algorithms, which are the core engine of this study, (3) Personalization Interface to deal with information filtering, and (4) Dual Representation of the semantic user profile.

The second component is devoted to Cluster Analysis in support of a personalized search for E-learning, which is an area of interest for both Data Mining and Information Retrieval. Cluster Analysis is used to divide the documents into an optimal categorization that is not influenced by the hand-made taxonomy of the colleges and course titles. In other words, clustering is used to both refine the college-based ontology and also as a mechanism to “shake” the rigidness of an otherwise entirely manually-constructed ontology that may not be appropriate for all users and for all times. The most important advantage of clustering from the personalization perspective is that the clusters are later used as automatically constructed labels for each user profile. Hence, depending on the document collection and its evolution, both the user profiles and their underlying ontology labels are allowed to change or evolve accordingly.

The third component demonstrates the HyperManyMedia platform as a recommender system where the learners can either apply User Relevance Feedback to filter information or Collaborative Filtering techniques as a community-based approach. The semantic retrieval system provides the user with user-relevance feedback by employ-
ing a variant of Rocchio’s Algorithm. In collaborative filtering, two methods were applied on HyperManyMedia: (1) $k$-nearest neighbors method, and (2) user-to-user based on fast XOR bit operations.

The fourth component is dedicated to Multilingual Semantic-based Information Retrieval (MLIR) that falls into Domain Specific Retrieval (E-learning). In this part, a synergistic approach between the Thesaurus-based Approach and the Corpus-based Approach was followed. The Thesaurus-based Approach brings more insight about the domain and the relationship between the concepts in the domain and presents them in a better formulated query; this helps the users to navigate the system in a way similar to a multilingual dictionary. The multilingual thesaurus can be considered as a bilingual ontology thesaurus and it organizes terms with respect to two languages (English and Spanish). We used a simple bilingual listing of terms, phrases, concepts, and sub-concepts. The hierarchical structure of the ontology is used to define the relationship between concepts/subconcepts. The Corpus-based Approach is considered as Term Vector Translation. A query translation method was used to retrieve multilingual documents with expansion techniques for phrasal translation.

The fifth component of our system is a Semantic Visualization method, which is presented as semantic networks. The formalization of the semantic graph has been built intuitively to solve a real problem which is browsing and searching for lectures in a vast repository of colleges/courses. This visualization combines Formal Concept Analysis (FCA) with Semantic Factoring to decompose a complex, vast concept into its primitive order to develop knowledge representation for the HyperManyMedia platform. This dissertation is implemented and evaluated on a real E-learning platform named HyperManyMedia.