CHAPTER 1

INTRODUCTION

“The question of whether computers will ever be “creative” in the sense that we speak of creative composing is rather similar to the problem of whether they “think.” Also we might ask: “What is meant by the term creative?” Being “creative” would seem to depend at the very minimum, like “thinking,” on having a computer operate on a self-sustaining basis, and to “learn from experience.” Moreover, it seems that what we first consider strokes of insight and manifestations of “creative thought” are, once they are analyzed and codified and particularly, codified to the extent that they can be processed by a computer, no longer “creative processes” in the usual sense.”

Machines Who Think (McCorduck, P., 2004)

The association between machines and thinking has been investigated for sometime, mainly in the field of Artificial Intelligence. This association started when Charles Babbage built his Analytical Engine, followed by Konrad Zuse with his creative work on the general-purpose computer, and Alan Turing’s Machine Intelligence [103]. In 1999, Tim Berners-Lee, with his invention of the Web, ignited this dream again with a prophecy of software agents that can act on behalf of humans, that can reason about data, in a place where the Web is being a machine-readable information whose meaning is well-defined by standards [19]. In 2002, a new infrastructure of the Web was defined to help
this dream come true. New ways of data representation, such as RDF \(^1\), were adopted to provide a common format for expressing information about information (metadata). However, in 2006, Tim Berners-Lee stated that “despite the progress of his dream (the Semantic Web), a lot of challenges still exist: the reuse of information is limited, how to effectively query an unbounded Web of linked information repositories? how to align and map different data models? how to visualize and navigate the huge connected graph of information? [20]” This dissertation is not going to answer all these questions, but it will put some spotlights on the importance of the Model Driven Architecture, which has relied on the concept that computers are unlike the Human mind, they do not have knowledge about how to present knowledge by themselves. Therefore Artificial Intelligence has introduced eight different commonly used types of Human knowledge and has tried to implement them on computers [53]: procedural knowledge, declarative knowledge, metaknowledge, heuristic knowledge, structural knowledge, inexact and uncertain knowledge, commonsense knowledge, and ontological knowledge. In this dissertation we focus on metaknowledge, heuristic knowledge and ontological knowledge. This dissertation will address the following questions: how we can (a) add Metadata to our resources to make them reachable, (b) extract the ontology of a specific domain to make it understandable from the semantic perspective, (c) re-model our system to be ready for the Semantic Web, (d) design a Personalized Semantic search model, (e) build a Cross-Language search model, (f) utilize User Relevance Feedback to provide a personalized filtering method to users, (g) use Collaborative Filtering techniques to take advantage of the similarity between users, and finally (h) map the ontology of the domain into a Visual search model.

Potentially, this research is going to investigate the design and implementation of multi-faceted search engine interface that does not rely on existing search tools that rely on a simple syntax-based description of the content or on keyword-centroid approach, it will express the contents in an ontology-based representation to achieve the semantic meaning of resources. This investigation will consider the most important difficulties

\(^{1}\)Resource Description Framework
in designing any search engine, such as designing a scalable data repository, re-scoring the semantic search engine based on extracting meaningful information and enabling semantic search in real-world information retrieval systems. In this context, different research questions will be addressed and evaluated based on the faceted model that we implement. We were faced with many challenges in the efforts to accomplish a full working vision of the semantics of our domain, but we obtained quite impressive results by using the standards (RDF and OWL), the framework (Jena ²) and the platform (HyperManyMedia ³).

1.1 General Scope of the dissertation

This dissertation falls within a new sector of the “Colliding Web Sciences” as represented in figure 1.1. The particular sector can be named “IR for E-learning.” This sector merges part of “Information Retrieval” in the context of the Web with part of Education, “E-learning.” It encapsulates E-learning with the areas of indexing and searching, query languages, text and multimedia languages, metadata, ontology modeling, and user profiles and personalization.

²http://www.w3.org/2001/sw/
³http://hypermanymedia.wku.edu: We proposed this term to refer to any educational material on the web (hyper) in a format that could be a multimedia format (image, audio, video, podcast, vodcast) or a text format (Webpage, PowerPoint).
Only minimal work has been done on proactively merging Information Retrieval (IR) with Education [1, 32] which is not directly related to the E-learning domain research. Why is there a need for a specific thrust, such as “IR for E-learning”? The answer is simple: In order to accurately respond to an online student query for a specific resource (a lecture) from a huge repository, to have a better understanding of online students’ behaviors using the IR system (such as a search engine) and to analyze the effect of the design and development of a new search engine for online students, we should tap the knowledge in educational pedagogy and how students, in general, and online students, in particular, use a search engine. This knowledge will help to have a meaningful understanding of their searching patterns. The need for this synergy arises from the fact that the number of online students has been growing significantly; nearly 20 percent 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; these figures show nearly a 10 percent increase over the number reported the previous year, and an increase of 9.7 percent for online enrollments that far exceeds the 1.5 percent growth of the overall higher education student population. These facts are based on a survey that represented the fifth annual report on the state of online learning.
in the U.S. higher education ⁴.

The fast growth of the Web ⁵ and the escalating number of people that seek knowledge and information has increased the popularity of search engines. “The WWW is a hypertext corpus of enormous complexity, and it continues to expand at a phenomenal rate, this can be viewed as an intricate form of populist hypermedia, in which millions of online participants, with diverse and often conflicting goals, are continuously creating hyperlinked content [86].” In 2004, a study about search engines [63] showed that the indexable Web was more than 11.5 billion pages. Google claimed to index more than 8 billion pages, MSN ⁶ about 5 billion pages, and Yahoo! ⁷ at least 4 billion pages. Nevertheless, Google, which was the largest search engine, had only 76% coverage of the Web and the refreshing process took several weeks to a month. This “information tsunami” of data, a tidal wave of unrelated, unorganized, uncontrolled information, pushed the Information Retrieval community harder to find a breakwater that can relate, organize, and control this information in a more efficient way. Conventional search engines were mostly text based. This has been effective for the first decade of the Web because most of the materials located on the Web were text based. Recently, however, the exponential increase of resources on the Web has shown that we need a different type of search engine, one that can adapt to the changes of Web 2.0 ⁸.

1.1.1 An Overview of A Desired Solution

A structured information retrieval model considers a user who is interested in specific information (document/lecture), under a specific topic, and with a specific media format. This user would recall that this specific information is under the “English” domain and that it contains information about a concept “Drama” and also recalls having watched these lectures in a video format. With a classical information retrieval model, this query

⁵http://www.zakon.org/robert/internet/timeline
⁶http://www.msn.com
⁷http://www.yahoo.com
⁸The term "Web 2.0" describes the changing trends in the use of World Wide Web technology and web design that aim to enhance creativity, communications, secure information sharing, collaboration and functionality of the web.
could be expressed as find pages where (Domain = “English,” Concept = “Drama,” Media = “Video”). As results, the system would retrieve all the video lectures under “English” with a specific concept which is “Drama.” In this particular case, the users might need to express their query through a sophisticated expression that they might find complex. They might be interested in an advanced search engine interface which allows them to run this query without explicitly writing it. The solution to this problem can be looked at from two sides:

- Designing a structured system that can deal with a hierarchical domain. This system can be used for browsing and searching. In this case, information can be organized in hierarchies of classes which group concepts under domains and classify this information in an ontology-like structure.

- Besides the ontology structure that governs the search and browsing tasks, the interface can be modified to facilitate the interaction between the user and the resources.

The interface could identify the domain that the user has recently visited and can show the semantic relation between the query that the user wrote and the concepts underneath it. This can be considered very effective in a large structured dataset. Showing the users the structure of the domain helps shorten the browsing time to find specific information. In addition, it could alert the user to the availability of other concepts akin to their interests. This allows the users to see the global context of the domain instead of just the information that they are seeking. The problem of any Web-based existing repository is that it is structured in sequential order. For example, under each domain in our repository, there are courses and under a course, there are sequential lectures. We assume that the users read/listen/watch these lectures in sequential order. In reality, this does not always happen. Users might want to navigate the system by clicking on hyperlinks and browsing from page-to-page to find specific information. For example, they may want to find all lectures related to “drama” no matter which course they are taking. In such a situation, a different organization of the repository is desired. However,
there is no point in reorganizing the whole repository in a new organization structure other than the way it is already organized. One way to accomplish such a goal is through the ontology recommendation structure. When a user queries for specific information, the retrieval system brings the resources related to this query and also shows the user all the concepts/subconcepts related to the submitted query. Building the ontology is done separately and does not affect the organization of the repository. The process of navigating through the ontology can be understood as a traversal of a directed graph. The linked nodes of the graph represent concepts that are semantically related. When the user hits a node in the ontology, the user can see the children nodes under that concept. If she is interested in navigating through the deeper level of the hierarchy, then she can continue navigating through these nodes, and at the moment they are interested in navigating in the opposite direction (upper concept), she can just click back to go to the parent node. In this way, the users are not restricted by the design of the system, and can freely navigate between concepts and relations that are represented by the ontology. The most essential problem is to build the most sophisticated, broad, domain specific ontology that can cover all concepts and their relationships. Not only that, this ontology should be dynamic, since the system is dynamic with the frequent adding of new courses, new lectures, etc.

The main objective of this research is to find empirical knowledge about the usefulness of using different methodologies to design an information retrieval system, either by using (a) metadata, (b) semantic knowledge representation, (c) natural language processing techniques, (d) clustering algorithms and (e) visual representation. In addition, we compare these models based on, scalability, efficiency, and usability. In chapter 3 (proposed approach: automated discovery, categorization and retrieval of personalized semantically enriched E-learning resources), we explain the architectural design of the following models: (a) metadata search model Section 3.1, (b) semantic search model using ontologies Section 3.2, (c) dual representation of the semantic user profile for personalized web search in an evolving domain Section 3.3, (d) augmented HyperManyMedia model Section 3.4, (e) hybrid recommender system Section 3.5, and (f) , Section 3.6 and
Consequently, we evaluate each section in Chapter 4 (Experimental Analysis).

1.2 Principal Concepts

1.2.1 HyperManyMedia

Today, an educational Website can be a gold-mining repository rather than a static Website—a voyage in space, time, and technology to discover the hidden student behavior and experience. In the department of Distance Learning at Western Kentucky University, an interactive Web environment was developed where teachers, researchers, and knowledge seekers can discover information about their distance learning students. Every single access to our platform, by any student to different types of learning material, such as text, audio, podcasting, and video lectures, was traced and recorded in log files. The audio and video lectures were presented through the latest technology, podcasting and vodcasting, to enhance the learning mobility. By tracking the behavior of each online student and knowing which lectures he/she has selected, the sequence of lectures that were selected, the type of the selection (text, audio, or video), and the method used (online or offline), we can build a user model (user profile), which is a system representation of how the learner relates to the conceptual structure of the application. Education has been (and still is) changing dramatically [42]. Several changes are occurring simultaneously, which are characterized by the three A’s: anyplace, anytime, anyhow. For the last few years, there has been a noticeable cultural trend among students who prefer to combine study and work. Online courses and the Internet play a major role in helping the students become independent of the physical availability of the teacher. However, online courses do not automatically make the teacher available anyplace and anytime. Some researchers have tried to create automatic teaching systems that simulate the intelligence of the teacher like SQL-Tutor [106] and ELM-ART [143] or Beal’s Artificial Intelligence inspired system [15]. “Intelligent Tutoring Systems” (ITS) can provide effective instruction, but learners do not always use such systems effectively. Motivated
students interested in course material take ITS readily, but others will improvise ways to get through without putting much effort.” [15]. As exposed, the ITS approach in general does not rely on the intelligence of the learner to select the information that he or she needs. In our approach, we allowed the students to choose the lectures that were needed, the sequence that was desired, and the learning style/format that was preferred. This approach provides answers to four A’s: Automatic Synchronization, Accessibility, Availability, and Adaptivity. Automatic Synchronization distinguishes both Podcasting and Vodcasting from the traditional multimedia (audio and video) on the Web. Most likely, Pod/Vodcasting will not replace traditional multimedia on the Web, but will rather become a more flexible extension of it, offering more diversity to a considerably larger audience. The key element of this intelligent technology is the automatic feed, which allows online students to subscribe to this feed only once, and the updated lectures, audio recordings of textbooks, texts, recent audio or video interviews, etc. to be automatically transferred to students’ MP3 devices. Accessibility means offering learners with different needs alternative ways to navigate through the information. For instance, the inclusion of closed caption text was embedded into our system to allow different ways of information delivery (hearing impaired students need the caption as an alternative to sound). Using this method, we can provide students with disabilities (hearing impaired) alternative ways of accessing online course materials. Availability means enabling online students access to lectures any way and anyhow (i.e., through the Internet via streaming the media online to a browser or an MP3 device or streaming the media offline which allows students to “read” or “review” texts while walking or driving). Adaptivity refers to learner preferences regarding different learning styles. Some learners prefer learning by reading (text), others by listening (audio) and yet others prefer a visual learning style (video). What is innovative about our system is that the four A’s can be encapsulated into the personalization aspect, which includes all aspects of the learning situation, such as personal preferences of student learning styles and needs. We called this system HyperManyMedia [166].
1.2.2 Information Retrieval (IR)

Information Retrieval (IR) is linked to the user information need which can be expressed as a query submitted to a search engine [10]. To accommodate this need, information has to be first analyzed and structured, then stored and organized in order to be retrieved.

**Definition** An information retrieval model is defined as “a quadruple \([D, Q, F, R(q_i, d_j)]\) where: \(D\) is a set of documents in a collection, \(Q\) is a set of queries or representations of the user information needs, \(F\) is a framework for modeling document representations, queries, and their relationships, and \(R(q_i, d_j)\) is a ranking function which associates a real number with a query \(q_i \in Q\) and a document representation \(d_j \in D\). such ranking defines an ordering among the documents with regard to the query \(q_i\).” [10]

Research has looked at an IR system from three perspectives: design, evaluation, and usage [87]. Regardless of which perspective, a system needs to be evaluated; this evaluation can be done using modeling, simulations, and user tests, with an emphasis on the user tests since no complete effective modeling can replace it. IR emerged in computer science in 1960 and continues to have a big impact in our daily activities, mostly through search engines, which have changed completely our way of searching, finding, or even validating information [87].

1.2.3 Personalization

Tailoring the user interaction with the Web is known as *Web Personalization*. This interaction varies from browsing simple to sophisticated websites, such as stock markets. The common goal is to provide the user with relevant information that fits his/her needs. To achieve effective personalization, we use different methodologies, such as data usage, click-stream data, data content, data structure, domain knowledge, and user demographic profiles, etc. Web personalization depends on efficient and intelligent techniques: Web mining, text mining, machine learning, clustering, and visualization to enhance the user’s web experience [6].
1.2.4 **Ontology (Knowledge Representation)**

The concept of ontology started being applied in the field of artificial intelligence as a mean for sharing knowledge [60]. Ontologies were defined as being “used for organizing knowledge in a structured way in many areas from philosophy to knowledge management and the Semantic Web” [40]. An ontology is described as a graph structure consisting of the following elements:

1. a set of concepts (vertices in a graph);
2. a set of relationships connecting the concepts (directed edges in a graph); and
3. a set of instances assigned to a particular concept (data records assigned to concepts or relation) [40].

Based on ontologies, Tim Berners-Lee, Ora Lassila, and Jim Hendler envisioned a new generation of the Web, called the Semantic Web [21]. The idea of the Semantic Web is to have a semantic representation of Web pages, people, and relations among them, thus providing us with more intelligent services [53].

1.2.5 **Natural Language Processing**

Natural Language Processing (NLP) which is also known as “Language Engineering” or “Language Technology” [101], is concerned with all those theories and hypotheses that deal with automatically processing textual information based on human knowledge of language, computational linguistics and speech language processing, etc., “What distinguishes language processing applications from other data processing systems is their use of knowledge of language [80].” NLP techniques are widely used and range from the very simple to the most complex including syntactics and semantic modeling.

1.2.6 **Clustering Documents/Text Analysis.**

“One application of clustering is the analysis of big text collections such as Web pages. The basic assumption, called cluster hypothesis, states that relevant documents tend to
be more similar to each other than to non-relevant ones. If this assumption holds for
a particular document collection, then the clustering of documents based on similarity
of their content may help to improve the search effectiveness.” [49] In particular, the
following improvements can be expected:

- Improving Search Recall: Search engines retrieve documents related to a specific
query term. Generally, the same concepts can be expressed using different terms,
thus searching for one of these terms will not retrieve the others. Clustering,
which groups together documents based on the overall similarity between them,
can improve the recall since the search query will match an entire cluster instead
of only one or more terms.

- Improving Search Precision: Assessing the relevance of documents to a query in a
big collection of documents is considered a difficult task. Clustering those docu-
ments into smaller collections, ordering them by relevance, and returning only the
most relevant group of documents, may help in finding a user’s specific interest.

A detailed analysis of different clustering algorithms is investigated in chapter 2.

1.3 Objectives and Contributions

1.3.1 Objectives

This dissertation targets the four objectives listed below.

- Objective 1: Metadata Driven Search Engine

  - Local Motivation:

    Western Kentucky University ⁹ hosts a HyperManyMedia open-source reposi-
tory of lectures ¹⁰. Thousands of online lectures are available in different
formats: text, powerpoint, audio, video, podcast, vodcast, and RSS. This
Web-based platform is the main medium of communication between WKU

⁹http://www.wku.edu
¹⁰http://hypermanymedia.wku.edu
online faculty and online students. For the last two years, the number of lectures added to this platform has grown significantly, and the usage of its resources by online students increased considerably. This growth has led to several problems:

1. **Searching for a specific college, course name, topic, or media format is time consuming, and the results are not always accurate.**

2. **Searching for combinations of results is impossible (e.g., finding all video lectures in the college of business related to accounting).**

- **Global Motivation:**

The fast growth of the Web and the escalating number of people that seek knowledge and information from it have increased the popularity of search engines. “The WWW is a hypertext corpus of enormous complexity, and it continues to expand at a phenomenal rate. The Web can be viewed as an intricate form of populist hypermedia, in which millions of online participants, with diverse and often conflicting goals, are continuously creating hyperlinked content”[86]. In 2004, a study about search engines [63] showed that the indexable Web was more than 11.5 billion pages. Google claimed to index more than 8 billion pages, MSN about 5 billion pages, and Yahoo at least 4 billion pages. Nevertheless, Google, which was the largest search engine, had only 76% coverage of the Web and the refreshing took several weeks to a month. The overwhelming engineering challenges mainly result from the philosophy of “one-size-fits-all”[31].

- **Objective:**

Our first objective is to design a focused metadata search engine to facilitate the search for learning objects for an open Web community (online students).

---

12[http://www.msn.com](http://www.msn.com)
13[http://www.yahoo.com](http://www.yahoo.com)
By “focused,” we mean that unlike a generic search engine, we will design a specialized engine that focuses on a limited collection of educational resources.

- **Objective 2: Personalized Semantic Information Retrieval System**

  - **Motivation:**
    
    Information filtering is concerned with the problem of delivering relevant information to the user. Typically, the relevance of information is related to the user preferences, which is called the *user profile* [17]. This led us to the notion of personalization. In an E-learning environment this personalization is based on context which requires intelligent methods for representing and matching both the learning *resources* and the variety of learning *contexts*. On the one hand, Semantic Web technologies can provide a representation of the learning content (lectures). On the other hand, the semantics of the user interests or profiles can form a good representation of the learning context that promises to enhance the results of retrieval via personalization. The key knowledge nugget in any personalization strategy for E-learning is an accurate user model. User Modeling is an active research area in information retrieval and personalization, especially when abstracting the user away from the problem [4], an abstraction that has, over the years, contributed to the design of more effective retrieval systems. Despite this improvement, the main focus in most information retrieval systems for the past decade has been on models that are “good for all users” [4], and not for a specific user. The enormous increase of information on the Web has 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, with some related work in [125, 55, 133, 99, 97, 100, 85, 26, 136].

  - **Objective:**
    
    The need for an infrastructure that can provide, manage, and collect data
that permits high levels of adaptability and relevance to the learner’s profile with innovative algorithms is a key requirement for the new generation of E-learning intelligent systems. Within this context, the Semantic Web is already a crucial factor that can contribute to the design of this infrastructure. Thus we follow a semantic Web-based approach to organize the E-learning domain and allow personalized adaptive search.

- **Objective 3: Cluster-based Information Retrieval System**

  - **Motivation:**
    While the college name which indicates an area of study seemed to offer a valid categorization basis, we thought that this might be too broad. For example, in the “math” college alone, the topics may vary from history of math to calculus, and geometry. A course-based categorization alone would also be obtuse because there may be courses with great overlap.

  - **Objective:**
    As a result, we propose to use a clustering technique to divide the documents into an optimal categorization that is not influenced by the hand-made taxonomy of the colleges and course titles. This can be expected to provide a finer granularity compared to the coarse college-based categories, and most importantly, promises to offer a more adaptive approach in the face of future additions of courses or colleges, or in the face of migrating the search approach to other E-learning platforms. In other words, clustering is used both to refine the college-based ontology constructed in Chapter 3 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.

- **Objective 4: Dual Representation of the Semantic User Profile for Personalized Web Search in an Evolving Domain**

  - **Motivation:**
    Several semantic-based user profile approaches have been introduced in the literature to learn about the user’s interests for personalized searching. However, many of them are ill-suited to cope with a domain of information that evolves and user interests that may change over time.

  - **Objective:** For this reason, we propose to handle the evolution of the domain and the learner profile by using a dual representation of the semantic user profile.

- **Objective 5: Augmenting the HyperManyMedia Repository with External Resources**

  - **Motivation:**
    The dataset of the WKU repository consisting of only 28 courses, a mounting to 2,812 learning objects (lecture), with each learning object represented in 7 different formats (text, powerpoint, streamed audio, streamed video, podcast, vodcast, RSS), thus a total of ~20,000 individual learning objects. Those materials are created by Western Kentucky University and located on the HyperManyMedia E-learning repository. We decided to extend these resources with external courses that are similar to our courses using an external open source repository. Those resources will be helpful to provide students with additional resources.

  - **Objective:**
    To extend the available resources, we provided WKU students with similar resources to their courses using external open source resources, from MIT
OpenCourseWare 14. Those resources are considered as pre-request or advanced resources.

- **Objective 6: HyperManyMedia As a Hybrid Recommender System**

  - **Motivation:**
    A recommender system, in an E-learning context, can be considered as an agent that tries to “intelligently” recommend actions to a learner based on the actions of previous learners. Such a recommender system could provide a recommendation to online learning materials or shortcuts. Those recommendations are based on previous learners’ activities or on the learning styles of the students that are discovered from their navigation patterns.

  - **Objective:**
    For this reason, we used User Relevance Feedback and Collaborative Filtering, where each user has been modeled based on his/her User Profile, and only the courses related to the user profile are presented semantically to the user. User Relevance Feedback provides the user with the capability to prune his/her profile by excluding documents he/she is not interested in. On the other hand, Collaborative Filtering is just the opposite, it extends a user’s profile based on the interests of communities of similar users.

- **Objective 7: Multi-Language (Cross-Language) Ontology-based Search Engine**

  - **Motivation:**
    A cross-language Search engine can be used to gain knowledge of understanding the same document in two different languages (e.g., English and Spanish). The following examples are not exhaustive but present the motivation behind adding Cross-Language ability to a search engine:

14http://ocw.mit.edu/OcwWeb/web/home/home/index.htm
1. A foreign or non-native speaker student who can understand a text written in English, but cannot formulate a good query to search for a document (lecture), so he/she can write a query in Spanish and retrieve the lecture in two languages (English and Spanish).

2. A native English speaker who would like to read the same document (lecture) in a foreign language (Spanish) to increase his/her foreign language knowledge. In this case, querying for a lecture in English and having the synonym in Spanish would provide the learner with the capability to pick the language he/she wants.

- **Objective:**

For the above reason, we present a multilingual course/lecture retrieval system. By multilingual, we mean that some courses are presented to students in two languages (English and Spanish). The exact lecture is presented in both languages (English and Spanish). When a user submits a query in any language (English or Spanish), if the query term exists in the corpus, the search engine will retrieve all the documents related to this query term and the ranking of the document will be based on Vector Space Model. All those documents will be from the same language the query term belong to. However, if the query term is part of the E-learning Ontology. The platform will retrieve the semantic meaning of this term; therefore, it will show all the classes/subclasses related to this query. In addition, it will show the translation of the query as a synonym in the alternative language. Therefore, when the user clicks on the translation of this query term, it will retrieve all the documents (lectures) related to that term and the ranking of these documents will be based on the Vector Space Model for this specific language.

- **Objective 8: Visual Ontology-based Search Engine**

  - **Motivation:**

A visual and interactive semantic representation of *HyperManyMedia* reposi-
tory can provide a summarization of the entire domain. This can be considered as a tool to visualize concepts and subconcepts. This visual exploration of documents enables users to have an overall view of the entire repository’s scope without even clicking on the resources and reading each document.

- **Objective:**

When a user submits a query to the visual search engine, the visual search engine matches the query to the whole visual ontology (concepts, subconcepts, etc) dynamically. Consequently, it presents all the sectors (concepts/subconcepts) that share some similarity with the query in different colors than the unmatched concepts. Therefore, the user can find what he/she is looking for immediately. However, as long as the user adds more letters in the search engine, the number of matched sectors narrow down to the most similar concepts in the ontology.

### 1.3.2 Contributions

- **Local Contribution:**

  Our contributions are:

  - **Design of a Metadata Search Engine** to overcome the limitations listed in Section 1.3.1. This search engine [29] is embedded on top of the HyperManyMedia platform at Western Kentucky University. This platform is now being used by WKU online students to search for specific formats of lectures more easily and more efficiently (details will be presented in Chapter 3).

  - **Design of a Personalized Semantic Search Engine** that adds another dimension to the metadata search engine that can manage and collect data that permits high levels of adaptability and relevance to learner’s profiles with innovative algorithms for the new generations of intelligent E-learning systems. Within this context, we designed a framework that consists of the following phases: (1) building the semantic E-learning domain using the known college
and course information as concepts and sub-concepts in a lecture ontology, (2) generating the semantic learners’ profiles (thus their user models) as an ontology from their navigation logs which record which lectures have been accessed, (3) clustering the documents to discover more refined sub-concepts (top terms in each cluster) than provided by the available college and course taxonomy, (4) re-ranking the learner’s search results based on the matching concepts in the learning content and the user profile, and (5) providing the learner with semantic recommendations during the search process in the form of terms from the closest matching clusters of the user profile. We decided to use a clustering technique to divide the documents into an optimal categorization that is not influenced by the hand-made taxonomy of the colleges and course titles. This can be expected to provide a finer granularity compared to the coarse college-based categories, and most importantly, promises to offer a more adaptive approach in the face of future additions of courses or colleges, or in the face of migrating the search approach to other E-learning platforms. In other words, clustering is used both to refine the college-based ontology constructed, refer to Section 3.2.2.1, 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 (details in Chapter 3).

– Design of a Dual Representation of the Semantic User Profile for Personalized Web Search in an Evolving Domain that provides a novel dual representation of a user’s semantic profile to deal with the problem of the user’s shifts of interests. This framework consists of adding the following phases to the previous design: (1) a lower-level semantic representation, con-
sisting of an accumulated gathering of user activities over a long period of time that uses a standard machine learning algorithm to detect user convergence, and (2) a higher-level semantic representation that detects shifts in the user activities; once this shift is detected, the higher-level semantic representation automatically updates the user profiles and re-initializes the system (details will be presented in chapter 3).

– Design of an Augmented *HyperManyMedia* Repository with External Resources that provides external resources to our users that are similar to our courses from MIT OpenCourseWare. It provides the students with additional (pre-request, advanced) resources (details will be presented in chapter 3).

– Design of HyperManyMedia As a Hybrid Recommender that uses User Relevance Feedback and Collaborative Filtering. Each user has been modeled based on interests (*User Profile*). Only the courses related to the user profile are presented semantically to the user. *User Relevance Feedback* provides the user with the capability to prune his/her profile by excluding documents that he/she is not interested in. On the other hand, *Collaborative Filtering* is just the opposite since it extends the user’s profile based on communities of similar users (details will be presented in chapter 3).

– Design of a Multi-Language Ontology-based Search Engine that presents a multilingual course/lecture retrieval system. By multilingual we mean that some courses are presented to students in two languages (English and Spanish). Our corpus consisting of courses/lectures from WKU (English only) augmented with courses from (MIT Open Courseware). The MIT courses contain parallel corpus lectures: the exact lecture presented in both languages (English and Spanish). When a user submits a query in any language (English or Spanish), if the query term exists in the corpus, the search engine will retrieve all the documents related to this query term and the ranking of the document will be based on Vector Space Model (details will be presented in chapter 3).
- **Design of a Visual Ontology-based Search Engine** that provides a visual knowledge representation of the *HyperManyMedia* ontology which is presented in the form of semantic networks. The formalization of the semantic graph intuitively has been built to solve a real problem which is browsing and searching for lectures in a vast repository of colleges/course. It combines *Formal Concept Analysis* with *Semantic Factoring* to decompose a complex, vast concept into its primitives in order to develop a knowledge representation for the *HyperManyMedia* platform. Our domain is considered as a graph, which represents the visual integrated ontology in the *HyperManyMedia* platform (details will be presented in chapter 3).

- **Global Contribution.**

As of 2006, the *HyperManyMedia* search engine has been ranked number 24 on “The Ultimate Guide to Using Open Courseware.” 15 (between Cambridge University and Harvard Business). The next five are Princeton, Stanford, Yale, Johns Hopkins, and Boston College. We consider this external recognition as an objective validation of what has been accomplished so far.

### 1.4 Dissertation Structure

The rest of this dissertation is divided into four chapters as follows:

**Chapter 2 (Literature Review):** We gave an overview of information retrieval models, personalization, semantic web, natural language processing, and clustering algorithms.

**Chapter 3 (Proposed Approach):** This chapter presents the core of this dissertation starting with the proposed architecture, the core algorithms, and the personalized interface. in addition, it presents a personalized semantic cluster-based search model, a hybrid recommender system, a cross-language search model and finally, a visualized ontology-based search model.

---

Chapter 4 (Experimental Analysis): In this chapter, the evaluation methods and experimental analysis are presented for each model implemented in Chapter 3.

Chapter 5 (Summary and Outlook): This final chapter summarizes the key findings of the experimental analysis in each model, then it compares these models to form conclusions, and moves to a discussion of the main objectives to design and implement this research and the benefit gained from using different intelligent structures and techniques before concluding with a brief outlook into future research.

The dissertation ends with the Appendices (A, B, C, ..., J), Index, Glossary, Bibliography and Curriculum Vitae.