A synergistic strategy for combining thesaurus-based and corpus-based approaches in building ontology for multilingual search engines

Leyla Zhuhadar¹

Dept. of Computer Information Systems, Western Kentucky University, Bowling Green, KY 42101, USA

ARTICLE INFO

Article history:

Available online 30 March 2015

Keywords:

Multi-language search engines
Cross-language search engines
Ontologies
Social Networks
Ontology co-learning

ABSTRACT

In this article we illustrate a methodology for building cross-language search engine. A synergistic approach between *thesaurus-based approach* and *corpus-based approach* is proposed. First, a bilingual ontology thesaurus is designed with respect to two languages: English and Spanish, where a simple bilingual listing of terms, phrases, concepts, and subconcepts is built. Second, *term vector translation* is used – a statistical multilingual text retrieval techniques that maps statistical information about term use between languages (Ontology co-learning). These techniques map sets of *t f i d f* term weights from one language to another. We also applied a *query translation method* to retrieve multilingual documents with an expansion technique for phrasal translation. Finally, we present our findings.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

In this article, we present a multilingual retrieval system. Our corpus consisted of courses/lectures from WKU (English only) augmented with courses from the MIT Open Courseware.² The MIT courses contain parallel corpora lectures (the exact lecture presented in both languages, English and Spanish). Our *MLIR* research falls into the *Domain Specific Retrieval*. The approach that we followed was a synergistic approach between (1) *Thesaurus-based Approach* and (2) *Corpus-based Approach*. In the case of the *Thesaurus-based Approach*, we used a simple bilingual listing of terms, phrases, concepts, and subconcepts. The hierarchical structure of the ontology is used to define the relationship between concepts/subconcepts. Also, we used a specific terminology that captures the domain of E-learning, those terms are associated with college name, course name, lecture name and presented in two languages. In the case of *Corpus-based Approach*, we used a *Term Vector Translation* approach, where the goal was to map statistical information about term usage between languages using techniques, which map sets of *t f i d f* term weights from English to Spanish and vice-versa. This research has been implemented on a real platform called *HyperManyMedia*³ at Western Kentucky University.

2. Background and related work

In this section we review a couple of concepts that are considered as the building blocks for designing Cross-language search engines such as, recommender system, ontologies, Natural Language Processing, and multi-language information retrieval systems.

2.1. Recommender system

One of the most powerful modes of personalization comes in the form of recommender systems (Nasraoui, 2005). Recommendation systems started back during the information retrieval era (McGill & Salton, 1983), but around the 1990s, it emerged as an independent research field (Adomavicius & Tuzhilin, 2005). The field of recommender systems can be classified into the following categories, based on how recommendations are made:

- Content-based: the user is recommended items (Web pages) based on his/her past activities (interest).
- Collaborative filtering: the user is recommended items (Web pages) based on people with similar interests in the past.
- Rule-based: the user is recommended items (Web pages) based on rules that enable limiting the recommended items to those that adhere to particular conditions.
- Hybrid-based: this model uses methods that combine from the above models, thus trying to avoid certain limitations in each one of the separate models.

¹ Ph.D. in Computer Engineering and Computer Science.

E-mail address: Leyla.zhuhadar@wku.edu

² <http://ocw.mit.edu/OcwWeb/web/home/home/index.htm>.

³ HyperManyMedia platform: <http://hypermanymedia.wku.edu>.

A recommender system in an E-learning context is a software agent that tries to “intelligently” recommend actions to a learner based on the actions of previous learners (Zaiane, 2002). 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. There are several approaches to automatically generate Web recommendations based on a user’s browsing patterns or explicit ratings (Nasraoui, 2005). Some rely on learning a usage model from Web access data or user ratings. For example, lazy user modeling is used in the most widespread form of *Collaborative Filtering* which stores all users’ information and then uses K-Nearest-Neighbors (KNN) to provide recommendations from the previous history of the K most similar users (Schafer, Konstan, & Riedl, 1999). Recently, others have used a different approach to recommend documents on the ground of the user profiles (de Gemmis, Semeraro, Lops, & Basile; Joachims, 2002; Joachims & Radlinski, 2007; Joachims et al., 2007). This approach learns from implicit feedback or past click history. Other ways to form a user model include using data mining, such as by mining association rules of the form: IF user views page A, THEN user views page B (Mobasher, Cooley, & Srivastava, 2000; Mobasher, Dai, Luo, & Nakagawa, 2001), or by partitioning a set of user sessions into clusters or groups of similar sessions. The latter groups are called session clusters or user profiles (Nasraoui, Krishnapuram, & Joshi, 1999). Even more recently, a Semantic Web usage mining methodology for mining evolving user profiles on dynamic Websites has been proposed (Nasraoui, Soliman, Saka, Badia, & Germain, 2008). This approach works by clustering the user sessions in each period and relating the user profiles of one period with those discovered in previous periods to detect profile evolution and also to understand what type of profile evolutions have occurred. This latter branch of using data mining techniques to discover user models from Web usage data is referred to as Web Usage Mining. A previous work on the use of Web mining for developing smart E-learning systems (Zaiane, 2002) integrated Web usage mining, where patterns were automatically discovered from users’ actions and then fed into a recommender system that could assist learners in their online learning activities by suggesting actions or resources to a user. A similar approach used hyperlink shortcuts by shortening frequent Web access sequences discovered in the Web log (Zheng, Niu, & Goebel, 2002). Another type of data mining in E-learning was performed on documents rather than on the students’ actions. This type of data mining is more akin to text mining (i.e., knowledge discovery from text data) than Web usage mining. This approach helps alleviate some of the problems in E-learning that are due to the volume of data that can be overwhelming for a learner. It works by organizing the articles and documents based on the topics and also providing summaries for documents.

The following section presents the most general algorithms to build a recommender search engine, then it discusses algorithms used by the most popular real recommender search engines, such as (1) Amazon.com and (2) Google Personalized News.

2.1.1. Building a recommender search engine

The design of a recommender search engine involves many different aspects; the following three are the most important:

1. **User Profile:** In general, this presents the user’s information, such as age, gender, interests, etc.
2. **Domain information:** It presents the type of information the domain provides to users, such as blog posts, documents (lectures), images, videos, tags, etc.

3. **User interactions:** The user’s type of interaction differs from domain to domain; it might be by user-to-user interaction of user-to-item interaction. The user might rate an item or a user’s content. The browsing behavior of users and the extracted patterns play a major role in deciding which type of recommendations will be offered to users.

However, the most important element in building a recommender search engine is the context component: (a) how the recommendations are presented to the user, (b) the ranking of these recommendations which is also known as *top item list*, (c) which categories/subcategories are considered as recommendations, which is also known as *personalized list of recommendations* for a specific user, etc.

2.1.2. Cases of recommender search engines in the real world

- **Amazon.com:** recommendations based on similar items (*items-based recommendations*). Amazon.com uses item-to-item collaborative filtering. The algorithm is described as follows: “*item-to-item collaborative filtering*: this method matches each of the user’s purchased and rated items to similar items, then combines those similar items into a recommendation list. To determine the most similar match for a given item, the algorithm builds a similar-items table by finding items that customers tend to purchase together (Linden, Smith, & York, 2003).”
- **Google News (news.google.com):** recommendations are based on the similarity between user profiles. This results in *user-based recommendations*, it is one of the most scalable recommender systems that provides personalized news for millions of subscribers. Google News uses three types of collaborative filtering techniques: (1) *MinHash Clustering*, (2) *Probabilistic Latent Semantic Indexing* (PLSI), and (3) *Covisitation counts*; more details in (Das, Datar, Garg, & Rajaram, 2007).

2.1.3. New research areas in information retrieval and search engine

Web 2.0 marked the beginning of the social media since users interact with each others through new social applications such as Facebook, MySpace, SecondLife, LinkedIn, Del.icio.us, Flickr, etc. Bruce Croft, in his latest book, “Search Engines: Information Retrieval in Practice, 2009 (Croft, Metzler, & Strohman, 2009),” distinguishes three areas of new research fields related to the searching/browsing mechanism. Croft defines this new Web paradigm as *Social Search*, which is “any application involving activities such as defining individual user profiles and interests, interacting with other users, and modifying the representations of the objects being searched (Croft et al., 2009).” An illustration of the emerging new areas of research in search is shown in Fig. 1. Croft explains each one of the four emerging areas of research in detail (Croft et al., 2009), the following sections provide a brief summary:

- **Filtering and recommendation:** As we mentioned previously in Section, one of the most important research trends in *information retrieval* is related to user-oriented search, such as personalization, user modeling, user relevance feedback, etc. In this section, we briefly summarize the new trends in research related solely to search engines. Filtering models in search engines have been divided into two categories (Croft et al., 2009): (1) static filtering models and (2) adaptive filtering models. A profile consists of “a single query, multiple queries, a set of documents, or some combination of these (Croft et al., 2009).” The static profile is generated one time and it cannot be changed over time, whereas the adaptive profile can be changed dynamically over time either

Fig. 1. Social search.

through a decision made by the user or automatically based on changes in the user's behavior. Croft indicates that the most common way of changing a user profile is through relevance feedback on documents (Croft et al., 2009).

- **Tag cloud search:** Starting from a user tagging an object, which could be a picture on Flickr, a video on Youtube, or a post on a Blog. The main difference in this type of process compared to the regular search engines is that instead of having the system index the terms automatically, the indexing is generated manually by the user, where each user specifies a term for an object. These are called User-generated ontologies (taxonomies) and referred to as *folksonomies*. (Croft et al., 2009) mentions three types of challenges in this type of research: (1) Since the tags are user-generated folksonomies, the tags are very sparse; therefore, the tag representation is complex. This is known as the *vocabulary mismatch problem*. Some solutions have been proposed such as stemming, pseudo-relevance feedback, and relevance modeling (Croft et al., 2009); (2) Tags are inherently noisy, like misspelling, spam, etc.; (3) Many objects in the collections are not tagged; Croft provides a solution to this problem called *inferring Missing Tags* (Croft et al., 2009).
- **Community search:** (Croft et al., 2009) described community search as *Searching within Communities*. This type of search is different from the traditional search since users are searching for either users or contents (Web pages, tags, etc.) related to their interests or hobbies. Croft distinguishes two types of community-based searching: (1) *community-based question answering (CQA)*, such as *Yahoo Groups*⁴ (this type of search engine uses retrieval models, such as *BM25* or *Language Modeling* to match questions and answers); (2) *Collaborative Searching* is divided into two categories: *Co-located Collaborative Searching* and *Remote Collaborative Searching*. The first type of search is allocated in the same location: Company, students, etc., whereas the second type is distributed over the world. An example of the Co-located Collaborative Searching is a search system named *CoSearch* (Amershi & Morris, 2008); another example of *Remote Collaborative Searching* is a search system named *SearchTogether* (Morris & Horvitz, 2007).

Recently, the development of reliable, scalable, and efficient community-base search engines gained considerable attention for both research players: industry and academia. However, new algorithms need to be designed to evaluate these new traits.

2.2. Ontologies

"An ontology is an explicit specification and formal specification of conceptualization of a domain of interest (Gruber, 1993)." The main goal of using an ontology in that work was to support sharing and reusing of formally represented knowledge in AI systems. To accomplish this, a common vocabulary need to be defined then used to represent the shared knowledge (Gruber, 1993). This included definitions of classes, functions, objects, and the relationships among all of them – which is an ontology. More specifically, the ontology represents the language of the Semantic Web. Since the Semantic Web will not replace the current Web, but will be built on top of it, a new structure was needed to deal with this issue. The old formal language, HTML needed to be preserved and a new semantic language needed to be used, the Resource Description Framework (RDF). RDF encapsulates the Web Ontology Language (OWL) in a schema similar to the XML format and it lays on top of it. Tim Berners-Lee proposed the following structure, illustrated in Fig. 2.

The proposed Semantic Web stack in Fig. 2 has been gradually refined. W3C⁵ provides updates of the current status of the Semantic Web.

What has been accomplished?

All of the following technologies are standardized URI, UNICODE, XML, RDF, RDFs and OWL:

- RDF (Resource Description Framework) is considered as the framework for the Semantic Web that allows the definition of triples.
- RDFs provide the vocabulary for RDF.
- OWL extends RDF and provides the description logic and the semantic reasoning.
- The query language is SPARQL.

What has not been yet accomplished?

The following technologies are not yet standardized:

- Trust layer (in progress).
- Digital Signature Layer (in progress).
- Rules (in progress).
- User Interface (in progress).

The Semantics level increases from the bottom layer of the Semantic Web stack (see Fig. 2) toward the upper layer. The

⁵ W3C: The main role of W3C is to guide us with the increasing levels of semantics and the related web technologies. In the following section, we briefly summarize the current state of Semantic Web: <http://www.w3.org/2001/sw/>.

⁴ <http://groups.yahoo.com/phrase/based-question-answering>.

Fig. 2. The semantic layers (Berners-Lee, Hendler, Lassila, et al., 2001).

relationship between these levels and ontologies have been mapped and expressed through the Ontology Spectrum, as illustrated in Fig. 3 (Daconta, Smith, & Obrst, 2003). Looking at this spectrum we can divide ontology levels into four distinguished categories: (1) Taxonomy, (2) Thesaurus, (3) Conceptual Model, and (4) Local Domain Theory. The semantic strength increases as we move from a lower category to an upper one. The main objective of ontology is making the knowledge reusable and shareable; thus, ontologies are constructed from vocabularies and their meanings. In this sense, we can compare this to the definition of an object in object-oriented programming languages. When we define an object, this object represents a class, and when we execute this program we create an *instance* of this class. Similarly, for ontologies, we have general concepts that represent *classes* and specific items that represent *instances*; we also have the *relationships*, *properties*, *functions* and *rules* among these concepts, etc. *Taxonomy* contains the structure of our domain represented as classes and subclasses, with the relationships between these classes/subclasses not defined in this level (weak semantics). A *Thesaurus* (RDFs) moves the ontology to a higher level where the associations and hierarchical relationships are defined. The *Conceptual Model* (OWL, UML, DAML, etc.) allows the definitions of class/subclass hierarchies. Finally, the *Local Domain Theory* (Modal Logic, First Order Logic, etc.) permits the software to understand data semantically with the highest level.

2.3. Natural Language Processing

Natural Language Processing (NLP) which is also known as “Language Engineering” or “Language Technology” (Manning, Schüt, & Press, 1999), 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.” (Jurafsky & Martin, 2008) NLP techniques are widely used and range from the very simple to the most complex including syntactic and semantic modeling. Language processing can be summarized into the following six criteria (Jurafsky & Martin, 2008):

- **Phonetics and Phonology:** the study of sounds (Sibawayhi, the Arabic grammarian of the 8th century was one of the first phonologist to study the vibration of sounds and words correlations (Edzard, 2000)).
- **Morphology:** the study of the meaningful components of the words.
- **Syntax:** the study of the relationships between words.

Fig. 3. The ontology spectrum (Daconta et al., 2003).

- **Semantics:** the study of meaning.
- **Pragmatics:** the study of the relationship between meanings in a speaking context.
- **Discourse:** the study of linguistics as a complete unit.

The main goal of this field is to enable human–machine translation, improving machine-machine communication, or simply processing languages in contextual and speech format. Our use of NLP concentrates on the *Semantics*.

2.3.1. Machine translation

The main idea of machine translation is to have a machine/software/agent capable of automatically translating a text or a speech from one language to another. Machine translation is a complex problem and it is far from being solved. Four different approaches to deal with machine translation can be distinguished (Manning et al., 1999):

- **Word-for-word approach**
- **Syntactic transfer approach**
- **Semantic transfer approach**
- **Interlingua approach**

The complexity of each approach increases top-down. First, starting from *Word-for-word approach*, which is considered as the simplest approach where each word is translated to an equivalent word. This approach is the simplest, but the most inaccurate and two major problems arise in this approach. One of them is the most common problem in NLP, ambiguity. Since there is no exact word-for-word translation and a lot of nuances in translating a word from one language to another language; this problem is considered a complex problem. The second problem is the order of words. The order of words differs from one language to another, and the meaning could be interpreted completely wrongly if the order did not follow the linguistics rules of each specific language (Manning et al., 1999).

The second approach is the *Syntactic approach*. In this approach the ordering problem we mentioned in *Word-for-word-approach* is solved, since this approach uses parsing rules that transfer the text from language to language. However, it does not solve the first problem in *Word-for-word-approach*. The third approach, the *Semantic approach*, depends on the semantic meaning of the text; the parser in this approach is more comprehensive, and it includes extra steps which present an intermediate step that encompasses the meaning of the text. This approach is better than the previous two, but it still faces problems that come from the nature of the language, “the literal meaning problem (Manning et al., 1999).” The last approach is the *Interlingual approach*. This approach uses knowledge representation that is independent from how the

language presents the meaning, it is considered as the best approach among the four. However, it is very difficult to design a thorough knowledge representation to present a language in a formalized manner. This is one of the biggest challenges in NLP. In the present research, the *Semantics transfer approach* in a very simplistic model is used.

2.4. Multi-Language Information Retrieval (MLIR)

Oard and Dorr (1996) defined *Multi-Language (Multilingual) Text Retrieval* as following: “The retrieval of documents or more precisely, electronic texts based on explicit queries formulated by humans using natural language, regardless of the language the documents and the query are expressed” (Fig. 4).

The majority of Information Retrieval systems are monolingual (English), even though only 6% of the world’s population has English as their native language (Haddouti, 1997). Surveys of Multi-Language Information Retrieval techniques and multilingual processing methods and application have been provided (Haddouti, 1997; Oard & Dorr, 1996). The major interesting reason for designing multilingual information retrieval systems can be summarized as follows:

A repository of documents written in multi-languages, with each individual document containing more than one language; for example:

- Technical documents written in non-English, but use expressions (jargon terms) written in English.
- A document that uses quotes written in languages different from the language of the article itself.
- A document that cites foreign articles and those citations are written in a language that is different than the language of the article itself.

The problem of a user who is capable of reading or using documents written in a specific language, but he/she is not fluent in this specific language to use the right query terms to find the document.

Three different scenarios to this problem are identified (Oard & Dorr, 1996):

- A user who is searching for images where those images are tagged and indexed in a language that the user does not understand.
- A researcher who is interested in a specific research topic and would like to know which individuals or institutes worldwide are working on the same topic.
- A user who has a system to translate documents to different languages and would like to search for those documents in languages with which he is unfamiliar.

Fig. 4. Machine translation mapped to cross-language search engine (Daconta et al., 2003).

On the other hand, in the first workshop on *Multi-Language Information Retrieval*, SIGIR96⁶ conference, the organizers divided the way of approaching the *Crossing Language Problem* into three approaches (Schauble & Sheridan, 1998):

1. Query translation.
2. Document translation.
3. Mix of query and document translation.

2.4.1. Approaches to Multi-Language IR

The research of MLIR can be divided into three approaches (Oard & Dorr, 1996):

1. Text translation Approach
2. Thesaurus-based Approach
3. Corpus-based Approach

2.4.1.1. Text translation Approach. A machine translation is used to map the query q and the document d into common language L . The difficulties of implementing such a system are explained in (Oard & Dorr, 1996), which also mentioned that the effectiveness of this approach is domain dependent—in some domains the quality is high and in others, it is very low. There were early implementations of the Text translation Approach (Davis & Dunning, 1995; Fluhr, 1997; Fluhr & Radwan, 1993), using straightforward techniques, but their main weakness is the low quality of the translation.

2.4.1.2. Thesaurus-based Approach. This approach is defined as an ontology-based approach (Oard & Dorr, 1996). Here, the thesaurus is an ontology, a knowledge representation of the domain. Four types of thesaurus are distinguished. One of the first implementations of this approach was the following two systems: (1) Salton augmented his SMART system to retrieve two languages (English and German). This was considered as the first MLIR system being tested and evaluated. Salton used Concept Lists, in the evaluation and average precision and there was different precision results for queries written in German versus English in this system; and (2) Pigur's system IRRD was based on a Vocabulary Thesaurus that used three languages (English, French and German); there was no evaluation tests for this system (Pigur, 1979).

2.4.1.3. Corpus-based Approach. Those techniques are exactly the same techniques used for monolingual information retrieval system. Instead of using a thesaurus, these techniques explore the statistical information about the corpus. Three techniques are distinguished (Oard & Dorr, 1996):

1. **Automatic thesaurus construction:** This approach extracts the statistical information about the terms in the corpus and automatically builds a thesaurus based on this information. For example, an algorithm to automatically extract terminology in bilingual corpus is used in (Pigur, 1979). Another algorithm finds noun phrase correspondences in a bilingual corpus. Another one used similar method to but it was based on linguistic knowledge (Daille, Gaussier, & Langé, 1994). The algorithm identified noun (N) phrases (F) in bilingual corpus (English and French), those NF most likely to be terms. Others extended previous models (Daille et al., 1994) by using word alignment and finding terminologies from bilingual corpus using a flow network model (Gaussier, 1998). Finally, another approach used a method based on the assumption that, probabilistically, there

⁶ SIGIR: Special Interest Group on Information Retrieval: <http://www.sigir.org/>.

is a correlation between the length of a text and its translation; the probabilistic score is applied to find the maximum likelihood of alignment of sentences (Gale & Church, 1991). More details about *Automatic thesaurus construction* is available in (Bruce, Metzler, & Strohmman, 2009; Grossman & Frieder, 2004; Oard & Dorr, 1996).

2. **Term vector translation:** This approach is defined as follows (Oard & Dorr, 1996): “We consider statistical multilingual text retrieval techniques in which the goal is to map statistical information about term use between languages... techniques which map sets of *t f i d f* term weights from one language to another.” (Oard & Dorr, 1996) Variations of techniques have been used to enhance the performance of this method (e.g., *relevance feedback*). For example, using query translation methods to retrieve multilingual documents (Davis & Dunning, 1995) where (Ballesteros & Croft, 1996) used *dictionary methods* for Multilingual information retrieval. Then they used an expansion technique for phrasal translation and query (Ballesteros & Croft, 1997). Finally, (Lavrenko, Choquette, & Croft, 2002) used a unified formal model using language modeling. They also integrated query expansion in addition to taking under consideration the most difficult problem in IR (disambiguaty). They implemented his model on both parallel corpus and dictionary.
3. **Latent Semantic Indexing (LSI):** In 1990, this technique was introduced in (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). It associates terms with documents based on the semantic structure in order to find the relevant documents to a query. This method is also used in MLIR (LSI-CL); for example, a system that retrieves documents in languages different than the query's language, in addition to the original language of the query, it uses LSI for French–English collection and the evaluation proved to have a good performance (Dumais, 2009). Another system, patented by Google, uses computerized Multi-language document retrieval using latent semantic indexing (Bruce et al., 2009; Feldman & Sanger, 2006; Grossman & Frieder, 2004; Ma, Pant, & Sheng, 2007; Oard, 1997).

2.4.2. New research areas in MLIR

The tremendous evolution of IR over the last decade gave rise to new research areas in *Cross-Language Information Retrieval (CLIR)*. Below are a few:

- *Interactive Cross-Language Retrieval (iCLR)*
- *Cross-Language Question Answering Retrieval (CLQAR)*
- *Cross-Language Image Retrieval (CLIR)*, *Cross-Language Video Retrieval (CLVR)*
- *Cross-Language Spoken Documents Retrieval (CLSDR)*

Over the last 13 years, *Multi-Language Information Retrieval (MLIR)* used different approaches, such as controlled vocabulary, dictionaries, thesauri, and free text. In general, MLIR relies on Machine Translation (MT), refers to Section 2.3.1. We should mention that one of the major contributors to the advances of MLIR is the *Cross-Language Evaluation Forum (CLEF)*. CLEF started in 2000, and “[it] promotes R&D in multilingual information access by:

- Developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and Multi-language contexts.
- Creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.

On one hand, the general research field of MLIR can be categorized into four major areas; these areas have been divided as follows (Peters, Braschler, & Gonzalo, 2003):

- **Multilingual retrieval:** In this field the IR system contains documents written in multiple languages and the goal is to query in one language and to be able to retrieve all the documents related to the query in multilingual format.
- **Bilingual retrieval:** In this system, the query is written in one language and the system is capable of retrieving documents in another language.
- **Monolingual retrieval:** The repository of this system contains documents in multiple languages; the IR system works as follows: when a user writes a query in one language, it will only retrieve the documents related to the query and the results are only from the same queried language.
- **Domain Specific Retrieval:** This research field is related to documents containing scientific text; the goal is to have an IR system capable of querying those terms in metalanguages and to retrieve documents in metalanguages.

On the other hand, the research of MLIR can be divided into three approaches (Oard, 1997): Text translation approach; Thesaurus-based Approach; and finally; and Corpus-based Approach.

3. Methodology

In this section, 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). Our corpus consist of courses/lectures from WKU presented in English language augmented with courses from the MIT Open Courseware⁷ that contains parallel corpora lectures (the exact lecture presented in both languages, English and Spanish).

Example 1. When a user submits a query in English or Spanish, if the query term exists in the corpora, the search engine retrieves all documents related to this query and ranks them based on the search engine ranking algorithm, all retrieved documents are in the language the query term belongs to. However, if the query term is part of the E-learning ontology (for more details about the design and implementation of our ontology, refer to our previous work (??) the system retrieves the semantic meaning of this term and it shows all the classes/subclasses related to this query, it also shows the translation of the query as synonym in the alternative language. When a user clicks on the translation of this query term, the search engine retrieves all documents (lectures) related to that term and rank them based on the search engine-ranking algorithm for this specific language.

3.1. MLIR Approach

Our MLIR research area falls into *Domain Specific Retrieval (E-learning)*. The approach we followed is a synergistic approach between (1) *Thesaurus-based Approach* and (2) *Corpus-based Approach*.

3.1.1. Thesaurus-based Approach

Thesaurus text retrieval allows the learners to explore more information during the searching process. The information retrieval system is capable of bringing more insight about the domain and the relationship between the concepts in the domain and present them in a better formulated query, this helps the learners navigate the system in a way similar to multilingual dictionary, but with visualized hints which can be considered as a powerful tool. Since we already designed and built a domain ontology, this

⁷ <http://ocw.mit.edu/0cwWeb/web/home/home/index.htm>.

Fig. 5. Cross-language search engine (English vs. Spanish).

part can be considered as an extension to the original ontology that can distinguish multilingual concepts/subconcepts and the relationship between the entities in the ontology.

A multilingual thesaurus can be considered as an ontology thesaurus. Therefore, a multilingual ontology is one which defines terms from more than one language. In our case, it is a bilingual ontology thesaurus, similar to dictionary, it organizes terms with respect to the two languages (English and Spanish). We used a simple bilingual listing of terms, phrases, concepts, and subconcepts. The hierarchical structure of the ontology is used to define the relationship between concepts/subconcepts. Since our ontology is a domain specific ontology (E-learning), the terminology used is not a standard terminology. We used a terminology that captures the domain, those terms are associated with college name, course name, lecture name and presented in two languages. Refer to the survey of multilingual text retrieval, by (Oard, 1998), for more details on Thesaurus types. In URL⁸ we present our complete extended Cross-Language E-learning Ontology ~40,000 line of code, Fig. 5 illustrates part of it. For more details on building ontologies, refer to (L. Zhuhadar & Kruk, 2010; Leyla Zhuhadar & Olfa Nasraoui, 2010; Zhuhadar & Kruk, 2010; Zhuhadar & Nasraoui, 2008; Zhuhadar, Nasraoui, & Wyatt, 2009; Zhuhadar, Rong, & Nasraoui, 2012; ZHUHADAR, NASRAOUI, & WYATT).

3.1.2. Corpus-based Approach

In Section 2, we reviewed different techniques to build a *Multilingual Information Retrieval* system some of these techniques explore the statistical information about the corpora. Oard and Dorr's survey (Oard & Dorr, 1996) distinguished three techniques: (1) *Automatic Thesaurus Construction*, (2) *Term Vector Translation*, and (3) *Latent Semantic Indexing (LSI)*.

Our approach is considered as *Term Vector Translation* (Oard & Dorr, 1996) defined this approach as: "statistical multilingual text retrieval techniques in which the goal is to map statistical information about term use between languages... techniques which map sets of *t fid f* term weights from one language to another."

We used a query translation method to retrieve multilingual documents with an expansion technique for phrasal translation. As we mentioned previously, our search engine uses the Vector Space Model to match the query term with the indexed documents, and it uses the scoring Eq. (2). The scoring algorithm is based on the vector space model representation of the documents. Each term vector representation is associated with each field document. We discussed the weight associated with each term in Section 3.1. We used the vector space model technique for *multilingual term vector translation*. Algorithm 2 describes the method used to implement this model. When a user submits a query in English or Spanish, and clicks on the cross-language search engine, if the query is a part of our indexed translated terms, the cross-language search engine does the following:

⁸ <http://people.wku.edu/leyla.zuhadar/researchSemantic/semanticnew.owl>.

1. Translate the query q to the alternative q' .
2. Use the vector space model to calculate the dot product between the translated query and the documents in the HyperManyMedia repository.
3. If the query has no translation in our system, then the user will have only the retrieved documents where terms from the original q query appears.

3.1.3. Synergistic approach between Thesaurus-based Approach and Corpus-based Approach

Our MLIR research falls into the *Domain Specific Retrieval (E-learning)*. The approach that we followed was a synergistic approach between (1) *Thesaurus-based Approach* and (2) *Corpus-based Approach*. In the case of *Thesaurus-based Approach*, we used a simple bilingual listing of terms, phrases, concepts, and subconcepts. The hierarchical structure of the ontology is used to define the relationship between concepts/subconcepts. Also, we used a specific terminology that captures the domain of E-learning; those terms are associated with college name, course name, and lecture name, and are presented in two languages. In the case of the *Corpus-based Approach*, we used the *Term Vector Translation* approach; the goal was to map statistical information about term usage between languages using techniques, which map sets of t fid f term weights from English to Spanish and vice-versa.

4. Evaluation

The design of the cross-language search engine followed a synergistic approach between a *Thesaurus-based Approach* and a *Corpus-based Approach*. Evaluating the Cross-Language Ontology-based Search Engine is based on the design we followed in Section 3, we considered the design as an extension to the original ontology that distinguishes multilingual concepts/subconcepts and the relationship between the entities in the ontology. More specifically, as a bilingual ontology thesaurus, similar to a dictionary, it organizes terms with respect to the two languages (English and Spanish). We presented the terminology that captures the *HyperManyMedia* domain; those terms are associated with college name, course name, and lecture name and were presented in two languages. We mapped the theory presented in Table 1 to a practical design of cross-language.

4.1. Research question

Will there be a difference in Top-n-Recall and Top-n-Precision when we Cross from the Spanish Language to the English language vs. from the English Language to the Spanish?

Table 1
Mapping theory into practice for multilingual thesaurus.

Type	Theory	Practical implementation
Subject thesaurus	Hierarchical representation of a domain with associative relationship between entities	College concept as a hierarchical upper-level
Concept list	Terms represents concepts in classes and subclasses	Subclasses represent the all the colleges in the HyperManyMedia domain
Term list	List of Multi-Language Synonyms	List of courses and lectures in cross-language synonyms
Lexicon	Semantics	Semantics represented in the OWL file to present the relations between all the upper types

Fig. 6. Top-n-Recall on multi-levels for cross-language search engine.

Fig. 7. Top-n-Precision on multi-levels for cross-language search engine (Cross-Language English to Spanish).

4.2. Evaluation results

We conclude that the cross-language search engine performs better when we cross from the Spanish language to the English language in the top-n-recall and Top-n-Precision, **which answers our research question** (Figs. 6 and 7).

We predict that the following reasons may have influenced the results:

- **English courses have been indexed and boosted in multiple stages during the design of the platform (during the last two years). Almost all the courses have been boosted by metadata tags and semantically enriched.**
- **Adding the Spanish courses was done during a very short period of time; thus we have not been able to add sophisticated tagging to these resources, because of the time constraints and the lack of understanding the language.**
- **The ontology relationships between the two languages need to be logically improved using a higher level of interrelationship between entities and concepts.**

5. Conclusion

In this article we illustrated a methodology for building cross-language search engine (ontology co-learning). A synergistic approach between *Thesaurus-based Approach* and *Corpus-based Approach* was proposed. First, a bilingual ontology thesaurus was designed with respect to two languages: English and Spanish. Second, a *Term Vector Translation* has been used. We also applied a query translation method to retrieve multilingual documents

with an expansion technique for phrasal translation. Finally, we present the evaluation results for this model. We found that the cross-language search engine performed better when we cross from the Spanish language to the English language in the top-n-recall and Top-n-Precision.

References

- Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749.
- Amershi, S., & Morris, M. R. (2008). CoSearch: a system for co-located collaborative web search.
- Ballesteros, L., & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval (pp. 84–91).
- Ballesteros, L., & Croft, B. (1996). Dictionary methods for cross-lingual information retrieval. *Lecture Notes in Computer Science*, 1134, 791–801.
- Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific American*, 284(5), 28–37.
- Bruce, W., Metzler, D., & Strohman, T. (2009). *Search engines: Information retrieval in practice*. Addison-Wesley.
- Croft, B., Metzler, D., & Strohman, T. (2009). *Search engines: Information retrieval in practice*. USA: Addison-Wesley Publishing Company.
- Daconta, M. C., Smith, K. T., & Obrst, L. J. (2003). *The semantic web: A guide to the future of XML, web services, and knowledge management*. New York, NY, USA: John Wiley & Sons, Inc.
- Daille, B., Gaussier, E., & Langé, J.M. (1994). Towards automatic extraction of monolingual and bilingual terminology (pp. 515–521).
- Das, A. S., Datar, M., Garg, A., & Rajaram, S. (2007). Google news personalization: Scalable online collaborative filtering (pp. 280).
- Davis, M. W., & Dunning, T. E. (1995). A TREC evaluation of query translation methods for multi-lingual text retrieval (pp. 483–498).
- de Gemmis, M., Semeraro, G., Lops, P., & Basile, P. A. Retrieval model for personalized searching relying on content-based user profiles.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- Dumais, S. T. (2009). An interdisciplinary perspective on information retrieval (pp. 1–2).
- Edzard, L. (2000). Sibawayhi observations on assimilatory processes and re-syllabification in the light of optimality theory. *Journal of Arabic Studies*, 3, 48–65.
- Feldman, R., & Sanger, J. (2006). *The text mining handbook*. Cambridge University Press.
- Fluhr, C. (1997). Multilingual information retrieval. *Cambridge Studies in Natural Language Processing Series*, 261–266.
- Fluhr, C., & Radwan, K. (1993). Fullex databases as lexical semantic knowledge for multilingual interrogation and machine translation (pp. 124).
- Gale, W. A., & Church, K. W. (1991). Identifying word correspondences in parallel texts (pp. 152–157).
- Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora (pp. 444–450).
- Grossman, D. A., & Frieder, O. (2004). *Information retrieval: Algorithms and heuristics*. Springer.
- Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5, 199.
- Haddouti, H. (1997). Survey: Multilingual text retrieval and access.
- Joachims, T. (2002). Optimizing search engines using clickthrough data. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 133–142.
- Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., & Gay, G. (2007). Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search.
- Joachims, T., & Radlinski, F. (2007). Search engines that learn from implicit feedback. *Computer*, 40(8), 34–40.
- Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing*. Prentice Hall.
- Lavrenko, V., Choquette, M., & Croft, W. B. (2002). Cross-lingual relevance models (pp. 175–182).
- Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 76–80.
- Ma, Z., Pant, G., & Sheng, O. R. L. (2007). Interest-based personalized search. *ACM Transactions on Information Systems*, 25(1), 5.
- Manning, C. D., Schüt, H., & Press, M. (1999). *Foundations of statistical natural language processing*. MIT Press.
- McGill, M. J., & Salton, G. (1983). *Introduction to modern information retrieval*. McGraw-Hill.
- Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on Web usage mining. *Communications of the ACM*, 43(8), 142–151.
- Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). Effective personalization based on association rule discovery from web usage data (pp. 15).
- Morris, M. R., & Horvitz, E. (2007). SearchTogether: an interface for collaborative web search (pp. 12).
- Nasraoui, O. (2005). *World wide web personalization*. Encyclopedia of data mining and data warehousing. Idea Group.
- Nasraoui, O., Krishnapuram, R., & Joshi, A. (1999). Mining web access logs using a fuzzy relational clustering algorithm based on a robust estimator. In *Eighth International World Wide Web Conference, Toronto, Canada*.
- Nasraoui, O., Soliman, M., Saka, E., Badia, A., & Germain, R. (2008). A web usage mining framework for mining evolving user profiles in dynamic web sites. *IEEE Transactions on Knowledge and Data Engineering*, 202–215.
- Oard, D. W. (1997). Cross-language text retrieval research in the USA.
- Oard, D. W. (1998). A comparative study of query and document translation for cross-language information retrieval. *Lecture Notes in Computer Science*, 472–483.
- Oard, D. W., & Dorr, B. J. (1996). A survey of multilingual text retrieval.
- Peters, C., Braschler, M., & Gonzalo, J. (2003). Advances in cross-language information retrieval: third workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 19–20, 2002: revised papers, Springer Verlag.
- Pigur, V. (1979). Multilanguage information-retrieval systems: Integration levels and language support. *Automatic Documentation and Mathematical Linguistics*, 13(1), 36–46.
- Schafer, B. J., Konstan, J. A., & Riedl, J. (1999). Recommender systems in E-commerce. In *Paper presented at the ACM conference on electronic commerce*.
- Schauble, P., & Sheridan, P. (1998). Cross-Language Information Retrieval (CLIR) track overview. NIST Special Publication SP (pp. 31–44).
- Zaiane, O. R. (2002). Building a recommender agent for e-learning systems. In *Proceedings of the International Conference on Computers in Education, 2002*, vol. 51, (pp. 55–59).
- Zheng, T., Niu, Y., & Goebel, R. (2002). WebFrame: In pursuit of computationally and cognitively efficient web mining. In *Paper presented at the PAKDD*.
- Zhuhadar, L., & Kruk, S. R. (2010). Intelligent social semantic collaborative filtering tools in an E-learning contexts. In *Paper presented at the IC-AI*.
- Zhuhadar, L., & Nasraoui, O. (2008). Semantic information retrieval for personalized E-learning. In *20th IEEE International Conference on Tools with Artificial Intelligence, ICTAI '08*, vol. 1 (pp. 364–368).
- Zhuhadar, L., & Nasraoui, O. (2010). Evaluating usability and precision of visual search engine. In *Paper presented at the proceedings of the 2010 spring simulation multiconference*, New York, NY, USA.
- Zhuhadar, L., Nasraoui, O., & Wyatt, R. (2009). Dual representation of the semantic user profile for personalized web search in an evolving domain. In *Paper presented at the proceedings of the AAAI 2009 spring symposium on social semantic Web, where Web 2.0 meets Web 3.0*.
- Zhuhadar, L., Rong, Y., & Nasraoui, O. (2012, 4–7 Dec. 2012). Toward the design of a recommender system: visual clustering and detecting community structure in a web usage network. In *Paper presented at the International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2012 IEEE/WIC/ACM.
- Zhuhadar, L., Nasraoui, O., & Wyatt, R. Model driven architecture: How to re-model an E-learning web-based system to be ready for the semantic web?
- Zhuhadar, L., & Nasraoui, O. (2010b). A hybrid recommender system guided by semantic user profiles for search in the E-learning domain. *Journal of Emerging Technologies in Web Intelligence*, 2(4), 272.