Query Expansion based on Domain Ontology for Learning Objects Search

Alejandra Segura N.
Dept. of Information Systems
Bio Bio University Concepcion, Chile
asegura@ubiobio.cl

Christian Vidal C.
Dept. of Information Systems
Bio Bio University Concepcion, Chile
cvidal@ubiobio.cl

Manuel Prieto M.
High School of Informatics
Castilla –La Mancha University Ciudad real, España
manuel.prieto@uclm.es

***Abstract*—This paper presents some research results which assesses the problem of query expansion in search activities. We propose the use of domain ontologies, linguistic processing and validation based on a general dictionary. The work is focused on Learning Object's search in specialized repositories. It includes a review of query expansion methods that can be used in e-learning and a description of a prototype that implement this approach. The results obtained so far allow us to state that the terms added to the query as result of the expansion process provide more specificity in the retrieved elements. Future research aims to demonstrate that with new terms, it is possible to improve relevance in retrieved Learning Objects.**

Keywords-component; Query expansion, Computer based instruction, Knowledge representation.

I. INTRODUCTION

The quantity of Learning Objects (LO) available in repositories or dispersed in the internet requires new techniques and more effectiveness to search and retrieve them.

The success in learning resources retrieval supposes; to express characteristics, restrictions or conditions about learning resource's use correctly, to select repositories or LO sources according to user's queries and, to present results that met user's expectations.

This work focuses on the first of these items, specifically in the formulation and the user's query processing. We expect to prove that through linguistic processing, the use of dictionaries and domain ontologies, the instructional designer's query terms become more specific. In later work, it will be evaluated if these processes provide more precision in LO retrieval.

The rest of this article is structured as follows. Section II, literature review section provides a related work overview about query expansions in learning object retrieval. Section III describes the process of query expansion and the prototype. The empirical results and some discussion are shown in the section fourth. Finally, the main conclusions are outlined in Section V, as well as the future work.

978-1-4244-5540-9/10/$26.00 ©2010 IEEE

II. QUERY EXPANSION

In general, most users typically formulate very short queries. Often they included terms that do not give sufficient information about the topic required〔1〕. To reduce the complexity associated with formulating queries, it is possible to use different strategies and techniques for reformulation, refinement or expansion.

The strategies and techniques can be classified according to interaction mechanisms used in the expansion process as interactive, automatics, or manuals〔2〕. In the interactive mechanisms, the user selects the relevant results and these are used as input in the expansion process. The automatic mechanisms do not have user intervention since the first results of the query are assumed as relevant results to perform the expansion process. Finally, the manual approach is characterized by the own user who refines the initial query according to the results and perform a new iteration〔3〕.

The query's ambiguity should be eliminated to increase the amount of relevant results. Therefore, it is necessary to focus on the query's context. The context is based on the knowledge about a particular domain or a specific task.

According to〔4〕, the query expansion can use methods such as lexical co-occurrence, clustering, stemming, and knowledge models. Lexical co-occurrence is the process to establish relationships among words based on the proximity analysis of the terms in a document. In clustering, documents that share a significant number of terms are grouped together in a cluster. The discriminant words from each cluster would be used for expanding the query. In this case, it is assumed that similar documents are relevant to the same queries〔5〕. The stemming is the process in which variations of terms are generated by the addition or removal of prefixes and suffixes as appropriate. This will extend or narrow the scope of the query. Finally, the knowledge-based methods extract the terms semantically related to user's query〔6〕.

The knowledge-based methods could be dependent or independent of the corpus. The corpus as a collection of text well constructed, balanced and annotated, where there the sense that the words take in a context is defined by the words that surround them.

Clustering and lexical co-occurrence methods obtain new terms from document collections. In contrast, based-knowledge model methods extract terms from models.

Follows several authors' opinions about the use of different query expansion methods.

Reference〔7〕proposed an hybrid technique based on semantic latency and query vectorial space. The expansion process is carried out with a matrix of terms/topics. This matrix indexes a set of relevant documents.

Semantic latency mechanisms were also used by〔3〕. They built a vector that represents full semantic of the query. Expansion process rather than comparing each query's word with the corpus' word uses a vector to compare the set of words with corpus〔3〕.

On the other hand,〔8〕integrated query expansion techniques based on corpus dependent knowledge models with lexical co-occurrence. The relevancy level of a specific term is determined by its co-occurrence with the general concept, which can be obtained from the corpus. The general concepts are substituted by a set of specific concepts used in the corpus that co-occurs with the user query's key concept.

Reference〔9〕proposed a search method based on the concepts instead of the strings. The concepts are dynamically ordered from the set of retrieved documents. More general terms are placed at a higher level followed by the specifics terms at a lower level. Relevant words are extracted from documents on top priority and are organized hierarchically using a subsumption function. This approach determines relationships between concepts and the way they are related.

Chli & De Wilde (2006) presented an interactive method for the expansion. It is based on the assumption that "documents contain terms with high information content that can summarise them and can be used to complement a general query string, to optimally reduce the search space for subsequent queries"〔4〕(pp.374). The documents can be summarized and used to complement a general query in order to optimally reduce the search space for subsequent queries.

Even though the large amount of expansion methods, query expansion is not always applicable. However, there is general agreement that it is appropriate when having short queries.

Reference〔5〕stated that there are other decisions that significantly affect the results and performance of the query expansion. Some of these decisions are related to the number of relevant documents, the way to select words or phrases to be used in the expansion, the weight and number of new terms, the exclusion of original terms, and whether the collections should be dependent on the domain. Specifically,〔10〕demonstrated that the number of new terms is less important than type and quality of the terms chosen.

A. Semantic Expansion having LO repositories searching

Regarding the search in LO repositories, there are other factors to be added to the previously presented issues.

LO are resources for learning, therefore, the search process should also consider educational requirements and constraints. These factors influence the selection of most appropriate and relevant LO to the user.

Reference〔11〕indicated that to reach a real level of reutilization of the Learning Object it is necessary to give them a semantic content that facilitates the search, selection, and composition.

The metadata's standards are mainly descriptive, that is to say not provide information of the meaning or semantic associated to them〔12〕.

When metadata are described based on dictionaries or ontologies, the semantic is explicit and, therefore, understandable and accessible to computers, other systems or software agents〔13〕.

Through the ontologies it is possible to structure, to homogenize and to give meaning to the metadata〔14,15〕and it is possible to enrich the query semantically and to rank the results according to your meaning and the related concepts〔11〕.

In this research, it is assumed that the user is an instructional designer. He seeks objects to be integrated in design of a course with the intention that students achieve the expected learning. Thus, this user must have a sufficient level of knowledge on the subject or discipline of the course.

As has been demonstrated in〔10〕, it is possible to support query formulations using knowledge models. They provide terminology and information about equivalence and hierarchical relationships between terms. This is beneficial not only for inexperienced users but also when they are expert in domain.

Most research related to ontology-based query expansion is focused on information retrieval in general and very few studies have been applied to the search for LO's in repositories.

One of the most recent works in this area was published by〔16〕. This study performs query expansion through the Java programming domain ontology (Java Learning Object Ontology - JLOO) defined by〔17〕. First, determine base concepts, i.e. these correspond to the ontology concepts that match initial query terms. Then, determine the user intention, which is to find out the subtree of ontology that represents the intention of user. To accomplish this, total impact of the base-concepts and related concepts in the ontology through synonym relationships is calculated. Finally, adding terms, only the words that belong to subtree of the user's intention must be added, except those that generate ambiguity, such as terms that are repeated in different subtrees of the ontology.

In the study's experiments, it is used a part of the JLOO ontology. Search is performed in the open web, not in LO Repositories. Also a set of relevant documents was extracted manually from the first ten pages retrieved. Considering these issues, the proposal can be efficient and accurate in retrieval of digital resources on the Web.

III. QUERY EXPANSION WITH LEARNING OBJECTS

To verify that the designer's queries are more precise through linguistic processing and the expansion based on domain ontologies, we implement a prototype that includes the following activities.

- Linguistic pre-processing of the query includes the extraction of stopword and the stemming of every term.
- Validation in a general dictionary and aggregation synonyms. At the moment, it has been used Wordnet dictionary for the query in English.
- Validation in the domain. In this case, it has been used the validation in ontology domain.
- Expansion based on incorporation of synonyms from the dictionary.
- Query expansion based on domain ontology. The algorithm of the〔16〕was adapted for this research according to restrictions or shortcomings identified:
- The intension of the user's query is based on calculating number of concepts that match between the query and synonyms represented in the ontology. This amount is divided by the total synonymous.
- In this way, it is harms expansion in short queries, or when there is a lot of synonyms in the ontology.
- The algorithm can not be used with a different domain ontology. That is, algorithm is dependent on the JLOO ontology structure. Particularly depends of synonymy's relations of and the hierarchical levels or depth of the ontology. If ontology is less deep, it will harm expansion.
- The final leaves level, called facts, has a set called facts, have a set of synonyms through which expansion is performed. That is, when query is expanded, it adds data properties or annotations associated with nodes (classes, subclasses) that form the subtree of the user's intention.
- The concepts included in the ontology are not processed linguistically. Specifically some have suffixes, prefixes or they are in plural (for example array, interfaces, literal constants, etc.). This implies
- that are reduced opportunities for expansion when it is make matching between query's terms with those represented in the model.
- In the JLOO Ontology the use of is_a relations is not consistent. These relationships do not necessarily reflect hierarchical relationships (i.e. father and son) between the concepts modeled.

In relation to problems presented above, this research proposed the following improvements:

- Stopword removal and stemming for queries and ontology's concepts.
- The concepts weight was calculated based on the initial amount of query terms.
- The queries definition and evaluation of expansion results will be performed for three teachers. They have experience in teaching, but having different knowledge level in JAVA programming domain.
- Regarding the restrictions associated with the ontology, in this first stage of this investigation will expand considering is-a relation as a part-of relationship. Additionally it will use annotations synonymous

To implement the expansion prototype, Java was used as programming language and as framework. In addition, other Tools and APIs such as Apache Apache Lucene, SPARQL Query Language for RDF, Protégé, Jena, ARQ Query Processor, WordNet (version 2.2) and JWNL are analyzed and integrated.

Figure 1 depicts an example in which the initial query is

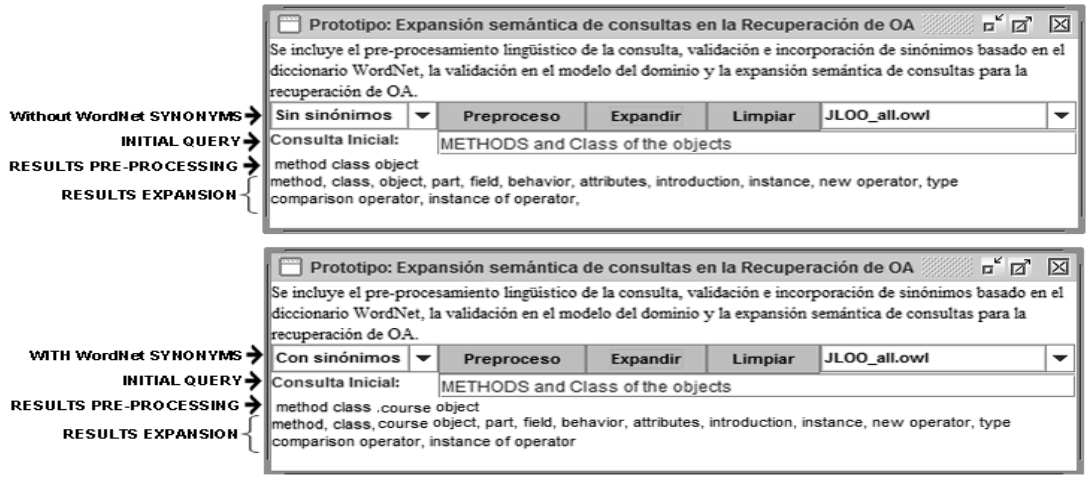

Figure 1. Pre-processing and Expansion example.

"Methods and Class of the objects". The prototype, first transforms the text to lowercase and removes the words "and" and "the" (stopwords). Then, it removes suffixes, prefixes or plural. In this case, the result is "class method object". To ensure that a word's stem is valid, it first checks whether it exists in the dictionary. But if a word is not in

dictionary, may be because it is badly written, or is a specific term of a domain. For example "else" "do while" or "arrays" are valid terms in the programming domain but not in general dictionary. Finally, there are terms that are valid in the domain without processing, for example, the words "programming", "for", "from", "testing" or "debugging" should not be stemming or eliminated as stopwords (for example for, from, else, etc.).

According to the above, the domain validation is performed before any processing. Since that the ontology is the domain model of the query, if a term is found in the ontology, it should go direct to the expansion process. Else, it will go first through the elimination of stopwords and stemming before being expanded.

IV. Results And Discussions

In the Learning Object Retrieval there are no standardized collections test data like those used in Information retrieval (IR) for example TREC- collection.

For this reason, in this research' stage, we are focused on proving that activities of processing, validation and expansion allows "add terms that refine or improve the specification of user's initial query". In future research will be necessary to define a mechanism to use metrics such as precision〔3,18〕.

In total, 16 queries were obtained. Table 1 shows the queries and the expansion results.

According to the experience and results, the following aspects are relevant.

- The main problem, related with knowledge models, is referred to correspondence that should exist between search terms and model concepts.
- The pre-processing linguistic of query terms is a core activity and indisputable in information retrieval. However, to ensure correspondence is necessary to consider too processing of ontology's concepts.
- The domain validation must be performed before any processing because there are terms that are valid only in specific domains.
- Due to queries specificity, the use of dictionary (e.g. Wordnet) for validation or synonym extraction does not give the expected results. In most cases only adds more ambiguity.
- Despite the problems or weaknesses identified, the linguistic processing and query expansion based on domain ontologies adds new terms that specify the query and therefore increase the chances of recovering more LO relevant for user.
- It is necessary to evaluate the effect and the relationship between the threshold of expansion and the structure of ontology. This is because the number of new terms depends on this relationship. No conclusive data regarding the amount of new terms, but most studies suggest an average of twenty〔5,19〕

TABLE I. Results of query expansion in programming JAVA domain

Initial Query	Expanded terms with JLOO's Ontology
Variables and types	variables, type, constants,
"Conditionals and recursion"	Conditionals and recursion,
methods	method, introduction, class, field, part,
"Fruitful methods"	Fruitful methods,
Expressions and types	expressions, type, statement,
Iteration and selection	iteration, selection, exception handling, sequence structure,
operators and expressions	operators, expressions, statement,
Object oriented programming	object, orient, program, attributes, behaviour, instance of operator, type comparison operator, new operator, instance, class,
"Abstract class"	abstract class, static binding, dynamic binding, abstract class structure, abstract method, polymorphism, overriding, overloading, superclass, subclass, inheritance, package, method prototype, interface structure, composition, interface, internal scope, external scope, scope,
Interfaces	interfaces,
"Constructor methods" Constructor	constructor, constructor method, copy constructor, class method use static, helper methods, void methods, method signature, method declaration, value returning methods, instance methods, field declaration,
Constants and attributes	constant, attributes, object, behaviour, instance of operator, type comparison operator, new operator, instance, class,
Arrays, Vectors	arrays, vector,
Lists	list,
Persistence	persistence,
"Structure control"	Structure control,

- Results, obtained using Lee's algorithm expansion, are dependent of the ontology depth. In tests with ontologies flat with lots of leaves, results are poor, because too many terms are added, causing an increase in ambiguity.
- Although today there are several domain ontologies, most of them do not have structure that allows them to be used with expansion algorithm proposed by [16].
- Successful expansion requires use of formal domain ontologies that are available for different areas of knowledge. Ontologies that are supported by the scientific community to ensure its continuous improvement. Therefore, every expansion algorithm should be based on ontological relations valid at any ontological model, independent of its structure, size or depth.

- Regarding success factors for effective use of ontologies in query expansion proposed by [错误!未找到引用源。] as found that both, the quality and user familiarity with the domain model influence positively in expansion results.
- For example, there are valid queries without expansion. This implies that ontology does not include all domain concepts or ontology concepts were not processed linguistically.
- Moreover, at the researching beginning, we assumed that the queries defined by users with more knowledge on the subject, would deliver better results. However, user familiarity with ontology domain was important in defining their consultations.
-

V. CONCLUSIONS

In the study's context, contribution obtained from a general dictionary such as Wordnet in query expansion is minimal because the teacher-user has knowledge in the domain and your queries are more specific that queries of other web user. It also justifies the importance of use represented knowledge in domain ontologies for query expansion in LO search.

Nowadays, most based-ontologies expansion algorithms use terminology relationships that are typical of a thesaurus rather than an ontological model. On the contrary, our next step in this research will concentrate on expansion procedures which privilege the ontological relations (eg. is-a or part-of relation) for extraction of new terms. To demonstrate the benefit of linguistic processing and ontology-based expansion in LO retrieval is necessary to design and implement a new experiment that prove the LO relevance through queries with or without expansion.

One aspect to consider in LO search, is that the teacher-user has knowledge on the subject and he must specify the educational requirements such as resource characteristics of learners, levels and interactivity's types that fosters, among others. This can be done based on the semantic contained in resources metadata and in search services of LO repositories. This should be done with the purpose of supporting novice users on issues related to instructional design.

ACKNOWLEDGMENTS

This work was partially supported by MECESUP UBB 0305 project, Chile; the TIN2007-67494 project of the Science and Innovation Ministry; The PEIC09-0196-3018 project of the Autonomous Government of Castilla-La Mancha.

REFERENCES

[1] A. Spink, D. Wolfram, M. Jansen, and T. Saracevic, "Searching the web: the public and their queries," Journal of the American Society for Information Science and Technology, vol. 52, no. 3, pp. 226–234, 2001.

[2] D. Harman, "Relevance feedback revisited," in SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, (New York, NY, USA), pp. 1–10, ACM, 1992.

[3] A. Abdelali, J. Cowie, and S. Soliman, Hamdy, "Improving query precision using semantic expansion," Information Processing and Management: an International Journal, vol. 43, no. 3, pp. 705–716, 2007.

[4] M. Chli and P. De Wilde, "Internet search: Subdivision-based interactive query expansion and the soft semantic web," Applied Soft Computing, vol. 6, no. 4, pp. 372–383, 2006.

[5] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," Information Processing and Management: an International Journal, vol. 43, no. 4, pp. 866–886, 2007.

[6] R. Navigli and P. Velardi, "An analysis of ontology-based query expansion strategies," in Workshop on Adaptive Text Extraction and Mining, Croatia, 22 September 2003 2003.

[7] L. A. F. Park and K. Ramamohanarao, "Hybrid pre-query term expansion using latent semantic analysis," in Data Mining, 2004. ICDM '04. Fourth IEEE International Conference on (K. Ramamohanarao, ed.), pp. 178–185, 2004.

[8] W. W. Chu, Z. Liu, and W. Mao, "Textual document indexing and retrieval via knowledge sources and data mining," Communication of the Institute of Information and Computing Machinery (CIICM), vol. 5, 2002.

[9] H. Joho, M. Sanderson, and M. Beaulieu, "A study of user interaction with a concept-based interactive query expansion support tool," in Advances in Information Retrieval, 26th European Conference on Information Retrieval, vol. 2997, (Sunderland, UK,), pp. 42–56, Springer Verlag, April 5-7 2004.

[10] A. Sihvonen and P. Vakkari, "Subject knowledge, thesaurus-assisted query expansion and search success," in In Proceedings of the RIAO 2004 Conference, (Paris: C.I.D.), pp. 393–404, 2004.

[11] E. Morales, A. Gil, and F. García, "Arquitectura para la recuperación de objetos de aprendizaje de calidad en repositorios distribuidos," in XII Jornadas de Ingeniería del Software y Bases de Datos, Zaragoza, 11 al 14 de septiembre de 2007 2007.

[12] J. Soto, E. Garcia-Barriocanal, and S. Sanchez-Alonso, "Semantic learning object repositories," International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL), vol. 17, pp. 432–446, 24 Marzo 2006 2007.

[13] G. Wenying and C. Deren, "Semantic approach for e-learning system," in IMSCCS '06: Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences - Volume 2 (IMSCCS'06), (Washington, DC, USA), pp. 442–446, IEEE Computer Society, 2006.

[14] K. H. Tsai, T. K. Chiu, M. C. Lee, and T. I. Wang, "A learning objects recommendation model based on the preference and ontological approaches," in ICALT '06: Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies, (Washington, DC, USA), pp. 36–40, IEEE Computer Society, 2006.

[15] M. Uschold, V. R. Benjamins, A. Gomez-perez, B. Ch, N. Guarino, and J. Robert, "A framework for understanding and classifying ontology applications," in in Proceedings of the IJCAI99 Workshop

on Ontologies and Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends. (IJCAI99)., 1999.

[16] [16] M.-C. Lee, K. H. Tsai, and T. I. Wang, “A practical ontology query expansion algorithm for semantic-aware learning objects retrieval,” Computers & Education, vol. 50, no. 4, pp. 1240–1257, 2008.

[17] M.-C. Lee, D. Yen Ye, and T. I. Wang, “Java learning object ontology,” in Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies, (Taiwan), pp. 538–542, IEEE Computer Society, 2005. 1078693538-542.

[18] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, “Evaluating implicit measures to improve web search,” ACM Transactions on Information Systems, vol. 23, no. 2, p. 22, 2005. Cited By (since 1996): 22Export Date: 14 December 2008Source: Scopus.

[19] M. Song, I.-Y. Song, X. Hu, and R. Allen, “Integration of association rules and ontologies for semantic query expansion,” Data & Knowledge Engineering, vol. 63, no. 1, pp. 63–75, 2007.