ISSN 1870-4069

An Algorithm for Semantic Expansion of Queries in a Boolean Information Retrieval System

Ana Laura Lezama, Mireya Tovar, David Pinto, Darnes Vilariño

Benemérita Universidad Autónoma de Puebla,
Faculty of Computer Science, Puebla,
Mexico

yumita1102@gmail.com, {mtovar,dpinto,darnes}@cs.buap.mx

Abstract. The increased amount of information in different domains, complicating the quick access to a particular need or specific query of any person or user, so arises the need to expedite this need, where the initial query is sought within a set of a domain documents chosen by the user. To retrieve more documents is important to incorporate techniques that allow to get more documents with the desired information. In this case extend the original query with the synonyms of the words that compose it can help improve the accuracy of the retrieval system (RS). In this paper, we propose an algorithm for query expansion of a Boolean Information Retrieval System (BIRS), in which the queries are formed by the concepts of four domain ontologies. According to the experimental results, we observe an improvement in the accuracy of the BIRS with the query expansion.

Keywords: Information retrieval system, semantic query expansion, ontology.

1 Introduction

Information Retrieval (IR) is the area of the science and the technology that tries to acquire, represent, and store information [1]. IR is the discovery of documents, usually unstructured documents (generally text) that satisfies a need for information within large stored collections in computers, through the development and implementation of intelligent techniques such as incorporating a retrieval information model, use of ontologies, etc [2].

Information Retrieval used to be an activity that only one person was dedicated to do it as a librarian, secretaries, etc., but with the exponential growth of information (text), makes it hard for a person to retrieve some information in a manner quick. Now the world changed and the information retrieval is performed through information retrieval models, training corpus, techniques of queries expansion, etc [2].

Information Retrieval System (IRS) are designed for processing text in natural language, rarely structured, and usually of semantic ambiguous. They allow

pp. 69–76; rec. 2016-07-18; acc. 2016-09-19

information retrieval, previously stored, through performing a series of queries. They have an information retrieval model and an inverted index [3].

One model of information retrieval is the Boolean model based on set theory and Boolean algebra. In this initial model, the user specifies a its query a boolean expression that it consists of a series of terms commonly linked by boolean operators such as *and, or* and *not.* Given the logical expression of the query, the system will return those documents that meet and form the set of relevant documents. Thus, the system partitions the collection of documents in two sets, those that meet the specified condition (relevant), and those that do not comply (not relevant). A document is therefore simply relevant or not. The popularity of the Boolean model, especially in the beginning, is given by its simplicity both conceptually, for the clarity of its formalism, as at the implementation level. Moreover, since queries are formulated as a Boolean expression, with semantic highly accurate, the user knows the reason why a document has been returned by the system, this does not always happen in other more complex models. Furthermore, since the documents are bag-of-terms, the recovery process is extremely fast [4].

In this paper, we present an algorithm to expand queries in a Boolean retrieval system that is used to find evidence of the concepts and semantic relationships of four domain ontologies in the corpus. The aim is to find evidence in the corpus of the concepts that exist in the ontologies, which are then used for the evaluation of them [5].

This paper is organized as follows: The most related works to query expansion in BIRS are discussed in section 2. In section 3, the proposed algorithm is explained. The experimental results and the data set are shown in section 4. Finally in section 5, the conclusion and future work are provided.

2 Related Work

Information Retrieval Systems with query expansion are systems that incorporate techniques commonly used in information retrieval to improve the documents retrieved by reformulation or by expansion of the original query, either by adding new terms or by weighting the query terms original. The terms can be automatically extracted from documents or taken from a linguistic resource [6].

In contrast the above, information retrieval systems without query expansion are those that process the query made by the user, only the information retrieval system is able to interpret, and recover the relevant documents to the query. But given that not perform query expansion by any technique, the system is only able to retrieve information associated with such query as it was entered, and as result of not add an expansion technique, IRS does not have the ability to access the most relevant documents to the user [7].

Some authors have used several techniques of queries expansion, linguistic tools and information retrieval models. Below, we expose some research related with this work.

In Chunyan et al. [8] propose a query expansion method based on Markov technique, discovering a relationship of terms through the concepts of a tree Markov model. For the query expansion, a given query and a set of documents, the authors calculate the probability. The results obtained in their experiments were effective.

Olufade et al. [9] proposes an improved method of document clustering for automatic query expansion tagged concept based thesaurus network was extended for the authors in other to obtain from an optimal cluster. Each query or terms is represented as a matrix of documents where each column describes a document and each row a query. The Fuzzy Latent Semantic Query Expansion Model process achieved a better precision and recall rate values on experimentation and evaluations when they compared with others existing information retrieval approaches.

De Campos et al. [10] given a document collection, the authors built a thesaurus based on a bayesian network, their method learn a poly-tree of terms. The poly-tree nodes represent terms of the collection in the form of binary variables. Given a query submitted to their system, the query expansion process starts placing the evidences in the learned polytree. This action means looking for the terms that appear in the query in the poly-tree nodes and setting their states to term more relevant. As their network is a poly-tree, they can use an exact and efficient inference method to propagate the probabilities. The results obtained from their experiments realized, show a improvement of the retrieval effectiveness using their query expansion technique proposal.

Lozada et al. [11] propose user relevance feedback, introduced two algorithms of the query expansion based on the function $VP{-}IDF$, build the first algorithm of query expansion, this algorithm receives as input the original query and submit the original query with it is expansion with weight for all terms of the query. Designate that the perfil user is composed by the number of the documents than the users has been evaluated, the number the documents relevant evaluated and a list the terms of user, in the that registry each term than appear in the documents evaluated by the user. The second algorithm named CE-DF is the that complete the query of user, with the terms more relevant profile and gives with results a list of terms in a text string, similarity to the gives for the user.

Mata et al. [12] propose an algorithm for Polish ontologies for task of information retrieval and for the query expansion, the authors used two types of cross references SeeRelatedDescriptor and ConsiderAlso and of the hierarchical structure of the ontologie MeSH, the expansion is produced to level term, and the terms that is descriptors in MeSH expands with the content of element SeeRelatedDescriptor or ConsiderAlso. SeeRelatedDescriptor associated the descriptor with others descriptors related through cross references, their object is provide others descriptors. ConsiderAlso reference to others descriptors related through linguistic roots. Their third strategy of expansion, is based in the tree structure as the structure of MESH for its descriptors. Also UMLS is very wide than MeSH, and it does use of concepts instead of terms and the authors developed two strategies of expansion, the first, is expand the concepts through relations between concepts and the second the expansion it is realized through the concepts relations for relations between concepts. The experiments realized demonstrated that the use of the hierarchical structure as the MESH was effectiveness.

In contrast these authors, in this article we present the expansion of queries using synonyms. They are extracted from WordNet, that are processed and incorporated in the SRIB, it is able to recover not only the documents that contain the original query, but also documents that contain synonyms of that query. According to the experimental results, an increase in terms of the information retrieved, compared with the system that does not perform expansion was obtained.

3 The Proposed Algorithm

In this section, we propose a general algorithm to expand the queries. They are initially formed by the words that make up to the concepts extracted from each domain ontology. Subsequently, the queries are expanded with the corresponding synonyms of each word, these are obtained from WordNet. The expanded queries are used by the Boolean Information Retrieval System (BIRS). The algorithm is described below.

1. For each domain ontology to extract concepts and semantic relationships.
2. For every word that form part of the concepts to extract the synonyms corresponding from WordNet [13].
3. Pre-processing domain corpus, concepts, relationships and synonyms. This step involves the following:
   (a) The reference corpus is split into sentences.
   (b) To remove special characters, punctuation, numbers and empty words.
   (c) To apply the Porter algorithm to the information [14].
4. To build queries. There are three types of queries:
   (a) Queries formed with the words of the concepts.
   (b) Queries formed with the synonyms of each word that integrate to the concepts.
   (c) Queries formed with the words of the concept that form the semantic relationship.
5. To apply the Boolean Information Retrieval System to the concepts, without expansion.
6. To apply the Boolean Information Retrieval System to the synonyms of the concepts, with expansion.
7. To mix and to join posting list, i. e., the results gotten with the BIRS with synonyms and without them.
8. To apply of the AND operator to the query that includes the two concepts that form the semantic relationship. The AND operator performs the intersection of the sentences that make up the posting of both concepts that form the semantic relationship.

In case of the evaluation of the results gotten, we use the Equations (1) and (2) for measuring the precision at the level of concepts and relationships:

$$P_C = \frac{Recovered\ concepts}{Total\ concepts}, \tag{1}$$

$$P_R = \frac{Recovered\ relationships}{Total\ relationships}, \tag{2}$$

where: *Recovered concepts* is the total of concepts obtained by the BIRS, and *Total concepts* is the number total of concepts in the domain ontology. In the case of recovered relationships, we evaluate for separated the taxonomic relationship and non-taxonomic relationship (for more information see [5]). The *Total relationships* correspond to the number total of relationships of each type in the domain ontology evaluated independently.

4 Results

In this section, we present the datasets used (4.1) and the results obtained in the experiments carried out (4.2).

4.1 Datasets

In Table 1 we present the number of concepts (C), taxonomic relations (T) and non-taxonomic relations (NT) of the ontology evaluated. The characteristics of its reference corpus are also given in the same Table: number of documents (D), number of tokens (T), vocabulary dimensionality (V), and the number of sentences (O). The domains used in the experiments are Artificial Intelligence (AI), e-learning (SCORM) [15], Oil (OIL) and Tourism (Tourism).

Table 1. Datasets.

Domain	Ontology			Reference corpus			
	C	T	NT	D	T	V	O
AI	276	205	61	8	11,370	1,510	475
SCORM	1,461	1,038	759	36	1,621	34,497	1,325
OIL	48	37	-	577	546,118	10,290,107	168,554
Tourism	963	1,016	-	1,801	877,519	32,931	36,505

4.2 Experimental Results

Below, we present the experimental results obtained by the algorithm developed and its comparison, i. e., the results of BIRS without query expansion and BIRS with query expansion. The results obtained by both BIRS in the algorithm,

Table 2. Results of the proposed algorithm for concepts.

	CO	C	F	P	CE	FE	PE	O	OE	DI	%
AI	276	274	2	0.992	274	2	0.992	1,994	3,332	1,338	67.10
SCORM	1,461	1,434	27	0.981	1,436	25	0.982	23,406	41,461	18,055	77.13
OIL	48	48	0	1.00	48	0	1.00	232,603	298,752	66,149	28.43
Tourism	963	682	281	0.708	788	175	0.818	86,353	249,087	162,734	188.45

for the case of the concepts, are shown in the Table 2 for each revised ontology (domain). We also show the total number of concepts extracted from the ontology (CO), concepts recovered by the BIRS without expansion (C), concepts that did not obtain associated sentences (F) and precision (P); concepts recovered by the BIRS with expansion (CE), concepts that failed to recover the SRIB with expansion (FE) and precision obtained (PE). In addition, the table shows the number of sentences obtained by the BIRS without expansion (O), with expansion (OE), the difference in the number of sentences recovered with expansion and without expansion (DI) and the percentage increase (%). In base to the results obtained for the concepts, we note that in the case of domains SCORM and Tourism, the number of concepts recovered is higher than the result of the BIRS without expansion. Furthermore, the number of sentences that contain the synonyms of the concepts increases the amount of sentences associated for each concept in the ontology, this occurs in each domain. The percentage increase of the information retrieved by the BIRS with expansion is greater than 28%, indicating that the concept can be represented in the corpus by its synonym corresponding and that this information is additional to that presented by the BIRS without expansion.

In Table 3 shows the results obtained by both Information Retrieval Systems with expansion and without it for taxonomic relationships. The RT column corresponds to the number of taxonomic relationships included in the corresponding domain ontology. The RR column is the number of recovered taxonomic relationships with BIRS without expansion and with BIRS with expansion (RRE). The F column shows the difference of relationships recovered by the BIRS without expansion and with expansion (FE). The precision of the system without expansion (P) and with expansion (PE). The amount of sentences recovered by the BIRS without expansion (O) and with expansion (OE) for this type of relationship, the difference obtained (DI) and the percentage difference (%) are also included. In base to the results obtained we observe that the number of relations taxonomic for two ontologies are maintained by the two algorithms designed. But in the case of SCORM ontology, the number of concepts is incremented by one, while for the Tourism ontology the number of concepts is increased from 291 to 441 this indicates that there are more concepts in the corpus that can only be found by its corresponding synonym. Also, the amount of sentences retrieved with the BIRS with expansion is increased for the four ontologies and even more for the ontology of Tourism, it support the existence of the synonyms for the concepts found in the corpus newly.

In the case of the non-taxonomic relationships, that only AI and SCORM

Table 3. Results of the proposed algorithm for taxonomic relationship.

	RT	RR	F	P	RRE	FE	PE	O	OE	DI	%
AI	205	205	0	1.00	205	0	1.00	782	972	190	24.29
SCORM	1,038	1,002	36	0.965	1,003	35	0.966	10,640	15,897	5,257	49.40
OIL	37	32	5	0.864	32	5	0.864	12,696	13,410	714	5.62
Tourism	1,016	291	725	0.286	441	575	0.434	5,606	22,552	16,946	302.283

Table 4. Results of the proposed algorithm for non-taxonomic relationship.

	RNT	R	F	P	RE	FE	PE	O	OE	DI	%
AI	61	61	0	1.000	61	0	1.000	108	136	28	25.92%
SCORM	759	738	21	0.972	739	20	0.973	8,728	10,155	1,427	16.40%

have it, we observe that the amount of recovered relationships is the same by both systems.

The RNT column corresponds to the number of non-taxonomic relationships included in the corresponding domain ontology. The R column is the number of recovered non-taxonomic relationships with BIRS without expansion and with BIRS with expansion (RE). The F column shows the difference of relationships recovered by the BIRS without expansion and with expansion (FE). The precision of the system without expansion (P) and with expansion (PE). The amount of sentences recovered by the BIRS without expansion (O) and with expansion (OE) for this type of relationship, the difference obtained (DI) and the percentage difference (%) are also included. In this case, some sentences were increased in the results with this type of relationship (see Table 4).

5 Conclusions

After experiments, with the proposed algorithm, we observe that the BIRS with expansion increase the number of sentences recovered for the domain ontologies. Even more, the number of concepts or relationship found are incremented in some case. Therefore, to expand the query with synonyms is a good alternative to get better precision of the system. The domain of Tourism had a satisfactory behavior with this algorithm, because its ontology has synonyms in the corpus that it's not possible recover with a traditional retrieval system.

As future work, we propose the use of lexico-syntactic patterns for the extraction of alternative synonyms from the text, synonyms than WordNet probably has not stored.

Acknowledgments. This research work has been partially supported by Sectoral Research Fund for Education CONACyT project 257357, VIEP-BUAP project 00570 and PRODEP-SEP project 00570 (EXB-792) DSA/103.5/15/10854.

References

1. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
2. de Campos, L.M., Fernández-Luna, J.M., Huete, J.F.: Query expansion in information retrieval systems using a bayesian network-based thesaurus. CoRR abs/1301.7364 (2013)
3. Kuna, H.D., Rey, M., Podkowa, L., Martini, E., Solonezen, L.: Expansión de consultas basada en ontologías para un sistema de recuperación de información. In: XVI Workshop de Investigadores en Ciencias de la Computación (2014)
4. Kwak, B.K., Kim, J.H., Lee, G., Seo, J.Y.: Corpus-based learning of compound noun indexing. In: Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 11. pp. 57–66. Association for Computational Linguistics (2000)
5. Lozada, C.A.C., Mendoza, E.E., Becerra, M.E.M., Flórez, L.C.G., Guzmán, E.L.: Algoritmos de expansión de consulta basados en una nueva función discreta de relevancia. Revista UIS Ingenierías 10(1) (2012)
6. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
7. Mata, J., Crespo, M., Maña, M.J.: Estudio del uso de ontologías para la expansión de consultas en recuperación de imágenes en el dominio biomédico. Procesamiento del lenguaje natural 47, 39–46 (2011)
8. Miller, G.A.: WordNet: A Lexical Database for English, vol. 38. ACM, New York, NY, USA (1995)
9. Onifade, O.F., Ibitoye, A.O.: Fuzzy latent semantic query expansion model for enhancing information retrieval. International Journal of Modern Education and Computer Science 2, 49–53 (2016)
10. Porter, M.F.: Readings in information retrieval. chap. An Algorithm for Suffix Stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997)
11. Tovar Vidal, M., Pinto Avendaño, D., Montes Rendón, A., González Serna, J.G., Vilariño Ayala, D.: Evaluation of ontological relations in corpora of restricted domain. Computación y Sistemas 19(1), 135–149 (2015)
12. Vechtomova, O., Wang, Y.: A study of the effect of term proximity on query expansion. Journal of Information Science 32(4), 324–333 (2006)
13. Vilares Ferro, J.: Aplicación del procesamiento del lenguaje natural en la recuperación de información en español. Ph.D. thesis, Universidad da Coruña, Departamento de Computación (Mayo 2005)
14. Yuan, C.: Concept tree based information retrieval model. Journal of Multimedia p. 652 (2014)
15. Zouaq, A., Gasevic, D., Hatala, M.: Linguistic patterns for information extraction in ontocmaps. In: Blomqvist, E., Gangemi, A., Hammar, K., del Carmen Suárez-Figueroa, M. (eds.) WOP. CEUR Workshop Proceedings, vol. 929. CEUR-WS.org (2012)