Fuzzy Information Retrieval Model Based on Multiple Related Ontologies

Maria Angelica A. Leite , Ivan L. M. Ricarte (Orientador)

Departamento de Engenharia de Computação e Automação Industrial (DCA)
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Caixa Postal 6101, 13083-970 – Campinas, SP, Brasil

{leite,ricarte}@dca.fee.unicamp.br

Abstract – With the World Wide Web popularity the information retrieval area has a new challenge intending to retrieve information resources by their meaning by using a knowledge base. Nowadays ontologies are being used to model knowledge bases. To deal with knowledge subjectivity and uncertainty fuzzy set theory techniques are employed. Preceding works encode a knowledge base using just one ontology. But a document collection can deal with different domain themes, expressed by distinct ontologies. In this work a way of knowledge organization and representation as multiple related ontologies was investigated and a method of query expansion was developed. The knowledge organization and the query expansion method were integrated in the fuzzy model for information retrieval based on mutiple related ontologies. The model performance was compared with another fuzzy-based approach for information retrieval and with the Apache Lucene search engine. In both cases the proposed model improves the precision and recall measures.

Keywords – Fuzzy Information Retrieval, Knowledge Representation, Query Expansion, Ontology.

1. Introduction

An information retrieval system stores and index documents such that when users express their information need in a query the system retrieves the related documents associating a score to each one. The higher the score the greater is the importance of the document [2]. Usually an information retrieval system returns large result sets and the users must spend considerable time until find items that are really relevant. Moreover, documents are retrieved when they contain the index terms specified in the queries. However, this approach will neglect other relevant documents that do not contain the index terms specified in the user's queries. When working with specific domain knowledge this problem can be overcome by incorporating a knowledge base which depicts the relationships between index terms into the existing information retrieval systems. To deal with the vagueness typical of human knowledge, the fuzzy set theory [13] can be used to manipulate the knowledge in the bases.

Knowledge bases in information retrieval cover a wide range of topics of which query expansion is one. The main aim of query expansion is to add new meaningful terms to the initial query. The expectation is that these new terms can improve the quality of retrieved documents bringing the most relevant and more semantically related to the initial query. A recent approach is to use ontologies to infer new terms to be added to the queries [3]. Usually information retrieval systems use just one conceptual structure to model the knowledge and compose the knowledge base [12, 14, 7, 17]. But the knowledge indexing a document collection can be expressed in multiple distinct domains. In some contexts these domains concepts are related by causal, spatial or similarity relationships. Each domain can be represented as a conceptual structure like a lightweight ontology. Lightweight ontologies include concepts, concepts taxonomies, relationships between concepts and properties that describes concepts [5]. The relationships between domains concepts can be translated to relationships between the lightweight ontologies concepts producing a knowledge base composed of multiple related lightweight ontologies. The use of the knowledge in a knowledge base composed of multiple related ontologies can be used to expand the user query with new concepts in order to improve information retrieval results.

In this work we focus on the fuzzy encoding of an information retrieval model which is supported by fuzzy related lightweight ontologies each one representing a distinct area of the knowledge domain. The model provides means to represent each ontology independently as well as their relationships. Based on the knowledge from the ontologies the system carries an automatic fuzzy query expansion. The documents are indexed by the concepts in the knowledge base allowing the retrieval

by their meaning. The results obtained with the proposed model are compared with the results obtained using just the user's entered keywords and with the results obtained by another fuzzy information retrieval system [4, 6]. The proposed expansion method is also employed in expanding queries for the Apache Lucene [1] search engine. The results show an enhance in precision for the same recall measures.

2. Fuzzy Information Retrieval Model

The knowledge is represented in multiple lightweight ontologies each one corresponding to a distinct domain. Each ontology representing a knowledge domain is a concept set $\mathrm{D}_k = \{c_{k1}, c_{k2}, \cdots c_{ky}\}$ where $1 \leq k \leq \mathrm{K}$, K is the domains number and $y = \|\mathrm{D}_k\|$ is the concepts number in each domain. These ontologies are related composing the model knowledge base. The concepts inside the ontology are organized as a taxonomy and are related by fuzzy specialization association (S) and fuzzy generalization association (G). Concepts pertaining to distinct ontologies are related by fuzzy positive association (P). Figure 1 shows the knowledge representation schema with multiple ontologies and the relationships between them.

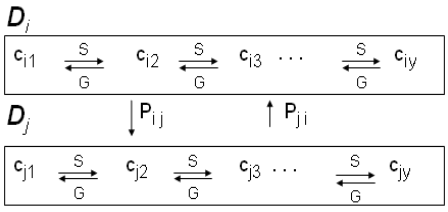

Figure 1. Knowledge representation with the relationships: fuzzy specialization (S), fuzzy generalization (G) and fuzzy positive assotiation (P).

The documents d_l from the collection are represented by the DOC set where $1 \leq l \leq \|DOC\|$. For each domain there is a relation that relates the documents from the DOC set to the concepts from the domain. This relation indicates the relevance of the concept to represent the document content. Its value is calculated following a *tf-idf* schema [2]. The query is expressed with the concepts from the distinct domains connected by logical operators like AND or OR.

In the proposed model the expansion method considers the knowledge expressed by the ontologies. The expansion is performed in two phases. In the first phase the fuzzy positive association (P) between concepts from distinct domains is used. In this step new concepts related to the ones present in the original query are added to the query. In the second step the fuzzy generalization (G) and the fuzzy specialization association (S) between concepts inside the ontologies are considered. In this step the more generic and the more specific concepts are added to the query.

Once the query is expanded the similarity between each document on the collection and the query is calculated. The similarity function associates a score to each document depending on how important the document is to the query. The documents are ordered in decreasing order of their score and are presented to the user as the final response.

3. Experimental Results

The experimental evaluation was carried out using a document collection sample referring to the agrometeorology domain in Brazil [10], a query set, a lightweight ontology referring to the geographical brazilian territory and a lightweight ontology referring to the climate distribution over the brazilian territory.

The ontologies were manually constructed considering a brazilian map [15] that contains the Köppen climate [16] distribution over the country. The first ontology refers to the brazilian territory, say domain D_1, with three levels. The root node is labeled 'Brazil', the descendant nodes are labeled with brazilian regions and each region node has the respective brazilian state nodes as descendants. Figure 2 shows a sample of the brazilian territory ontology. For the brazilian territory ontology the fuzzy generalization association and the fuzzy specialization association relates the spacial relationship between the territory entities they refer to. The second ontology refers to the climate distribution over the brazilian territory, say domain D_2. The root node is labeled 'Climate', the root descendant nodes are labeled with brazilian zonal climates and each zonal climate has the respective associated Köppen climate nodes as descendants. Figure 2 shows a sample of the brazilian climate ontology. For the brazilian climate ontology the fuzzy generalization

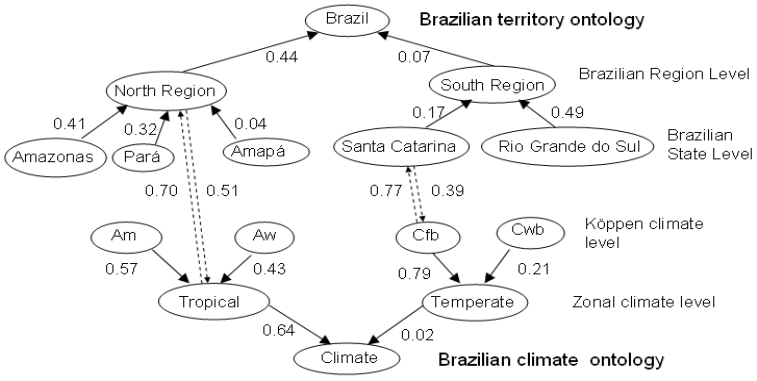

Figure 2. Brazilian territorial and brazilian climate lightweight ontologies and their fuzzy associations.

association and the fuzzy specialization association relates the spacial relationship between the climate entities they refer to. The fuzzy positive association between the ontologies is established by the distribution of the climate over the brazilian territory. The association is settled in two levels. The first one is between the brazilian regions and the zonal climates and the second one is between the brazilian states and the Köppen climates. The dashed lines in Fig. 2 illustrate both associations levels.

The experiment was ran to test the proposed model performance and to compare its performance with a similar approach, that is, the multi-relationship fuzzy concept network information retrieval model [11]. The experiment also tested the use of the query expansion method in the Apache Lucene text search engine [8]. All the models showed an improved behavior in precision and recall measures when compared with using just the keywords entered by the user. Considering a given query, recall is the ratio between the number of documents retrieved and considered relevant, and the total number of documents considered relevant in the collection and precision is the ratio between the number of documents retrieved and considered relevant, and the total number of retrieved documents [2]. The proposed model showed better results than other models considering the proposed knowledge representation, as multiple related ontologies, and the query expansion method that uses this knowledge organization. Another experiment tested the ontologies as crisp ones where the fuzzy generalization and the fuzzy specialization associations assume values in the set $\{0, 1\}$ denoting the existence (1) or absence (0) of the relationship between concepts [9]. Also, in this case, the proposed model obtained good performance considering the precision and recall measures.

4. Conclusion

The use of knowledge in the information retrieval process can enhance the quality of the results by returning documents semantically related to the initial user's query. To deal with the uncertainty and vagueness present in the knowledge the fuzzy set theory has been used. When working with specific document collections one way to accomplish that is to represent the knowledge in distinct domains each one being represented as an independent lightweight ontology. These domains knowledge can be related to each other composing a knowledge base. To explore these issues this work presented an approach for improving the document retrieval process. Contrary to other approaches that consider the knowledge base composed of just one ontology, the proposed model explores knowledge expressed in multiple ontologies whose relationships are expressed as fuzzy relations. This knowledge organization is used in a novel method to expand the user query and to index the documents in the collection.

Experimental results show that the proposed model achieves better performance when compared with other fuzzy information retrieval approach. The manually constructed knowledge base offers a semantic knowledge that leads to good retrieve performance. As the knowledge organization and representation as ontologies is a growing area, the model offers a way where these independently developed ontologies can be reused in the task of the information retrieval process.

References

[1] Apache Project. Apache Lucene Overview. Página na internet, The Apache Software Foundation, Acesso em: Agosto 2008. http://lucene.apache.org/java/docs/index.html.

[2] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[3] J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing and Management*, 43(4):866–886, 2007.

[4] Shyi-Ming Chen, Yih-Jen Horng, and Chia-Hoang Lee. Fuzzy information retrieval based on multi-relationship fuzzy concept networks. *Fuzzy Sets and Systems*, 140(1):183–205, 2003.

[5] Asunción Gomez-Pérez, Mariano Fernández-Lopez, and Oscar Corcho. *Ontological Engineering*. Springer-Verlag, 2003.

[6] Yih-Jen Horng, Shyi-Ming Chen, and Chia-Hoang Lee. Automatically constructing multi-relationship fuzzy concept networks for document retrieval. *Applied Artificial Intelligence,*, 17(1):303–328, 2003.

[7] Raymond Y. K. Lau, Yuefeng Li, and Yue Xu. Mining fuzzy domain ontology from textual databases. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 156–162, Washington, DC, USA, 2007. IEEE Computer Society.

[8] Maria Angelica Leite and Ivan L. M. Ricarte. A framework for information retrieval based on fuzzy relations and multiple ontologies. In *IBERAMIA '08: Proceedings of the 11th Ibero-American conference on AI*, pages 292–301, Berlin, Heidelberg, 2008. Springer-Verlag.

[9] Maria Angelica Leite and Ivan L. M. Ricarte. Using multiple related ontologies in a fuzzy information retrieval model. In *Third Workshop on Ontologies and Their Applications*, Bahia, Brasil, 2008. Universidade Federal da Bahia.

[10] Maria Angelica A. Leite and Ivan L. M. Ricarte. Document retrieval using fuzzy related geographic ontologies. In *GIR '08: Proceeding of the 2nd international workshop on Geographic information retrieval*, pages 47–54, New York, NY, USA, 2008. ACM.

[11] Maria Angelica A. Leite and Ivan L. M. Ricarte. Fuzzy information retrieval model based on multiple related ontologies. In *20th IEEE International Conference on Tools with Artificial Intelligence*, pages 309–316, Washington, DC, USA, 2008. IEEE Computer Society.

[12] Christian Paz-Trillo, Renata Wassermann, and Paula P. Braga. An information retrieval application using ontologies. *Journal of the Brazilian Computer Society*, pages 17–31, Março 2006.

[13] Witold Pedrycz and Fernando Gomide. *Fuzzy Systems Engineering: Toward Human–Centric Computing*. John Wiley & Sons, Inc, 2007.

[14] Raquel Pereira, Ivan Ricarte, and Fernando Gomide. Fuzzy relational ontological model in information search systems. In *Elie Sanchez. (Org.). Fuzzy Logic and The Semantic Web*, pages 395–412, Amsterdan, 2006. Elsevier B. V.

[15] Projeto SISGA. Mapa do Clima no Brasil. Página na internet, Universidade Regional de Blumenau, Acesso em: Junho 2008. http://www2.inf.furb.br/sisga/educacao/ensino/mapaClima.php.

[16] Wikipédia. Classificação climática de Köppen-Geiger. Página na internet, Wikimedia Foundation, Acesso em: Junho 2008. http://pt.wikipedia.org/wiki/Classificação_do_clima_de_Köppen.

[17] Lei Zhang, Yong Yu, Jian Zhou, ChenXi Lin, and Yin Yang. An enhanced model for searching in semantic portals. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 453–462, New York, NY, USA, 2005. ACM.