Using Multiple Related Ontologies in a Fuzzy Information Retrieval Model

Maria Angelica A. Leite[1,2] and Ivan L. M. Ricarte[2]

[1] Embrapa Agriculture Informatics
PO Box: 6041 - ZIP: 13083-970 - Campinas - SP - Brazil
angelica@cnptia.embrapa.br
http://www.cnptia.embrapa.br
[2] School of Electrical and Computer Engineering, University of Campinas
PO Box 6101, Postal Code: 13083-970 - Campinas, SP, Brazil
{leite,ricarte}@dca.fee.unicamp.br
http://www.fee.unicamp.br/

Abstract. *With the Semantic Web progress many independently developed distinct domain ontologies have to be shared and reused by a variety of applications. The use of ontologies in information retrieval applications allows the retrieval of semantically related documents to an initial users' query. This work presents a fuzzy information retrieval model for improving the document retrieval process considering a knowledge base composed of multiple domain ontologies that are fuzzy related. Each ontology can be represented independently as well as their relationships. This knowledge organization is used in a novel method to expand the user initial query and to index the documents in the collection. Experimental results show that the proposed model presents better overall performance when compared with another fuzzy-based approach for information retrieval.*

Key words: Fuzzy information retrieval, ontology

1 Introduction

With the grown availability of information many research has been made in order to provide intelligent ways to easy the information access. One point is to treat not only the lexical information features but also to consider its semantics, that is, the meaning attached to it. Within this approach the usage of ontologies to organize the knowledge and to express semantic meaning has gaining attention. 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 the document relevance [1]. Usually 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. Knowledge bases can be manually developed by domain experts or automatically constructed from the knowledge in the document collection [2, 3]. To deal with vagueness typical of human knowledge, the fuzzy set theory can be used to manipulate the knowledge in the bases. It deals with the uncertainty that may be present in document and query representations as well as in their relationships. Knowledge bases in information retrieval cover a wide range of topics of which query expansion is one. A recent approach is to use ontologies to infer new terms to be added to the queries [4]. Usually information retrieval systems use just one conceptual structure to model the knowledge and compose the knowledge base. 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 domain's concepts can be translated to relationships between the lightweight ontologies concepts producing a knowledge base composed of multiple related lightweight ontologies. Consider the territorial division and the climate domains. These are distinct domains but the relation of a territorial division and a climate classification can be done through observation of geographic and climatic maps. The geographic domains are, in general, organized in an hierarchical way and can be represented by domain ontologies. Figure 1 shows the Köppen climate[3] distribution over brazilian territory[4] and Fig. 2 shows the ontologies referring to brazilian territory and climate domains respectively. The idea is to relate the ontologies by establishing fuzzy relationships between territory and climate concepts based on spatial distribution in the map. The dashed lines illustrates this kind of relationship.

We present an information retrieval model which is supported by fuzzy related lightweight ontologies each one representing a distinct knowledge domain. The model provides means to represent each ontology independently as well as their relationships. This way existent ontologies can be reused in the model. Based on the knowledge from the ontologies the system carries an automatic fuzzy query expansion. The documents are indexed by the concepts in the ontologies allowing the retrieval by their meaning. The documents do not need to be indexed by each ontology concepts as in a faceted approach. Given a query with concepts from an initial domain new semantically related documents indexed by other domain ontologies can be retrieved based on the ontologies relationships. The results obtained with the proposed information retrieval 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 [6, 7]. The proposed expansion method is also employed in expanding queries for the

[3] http://en.wikipedia.org/wiki/Koppen_climate_classification

[4] http://campeche.inf.furb.br/sisga/educacao/ensino/mapaClima.php

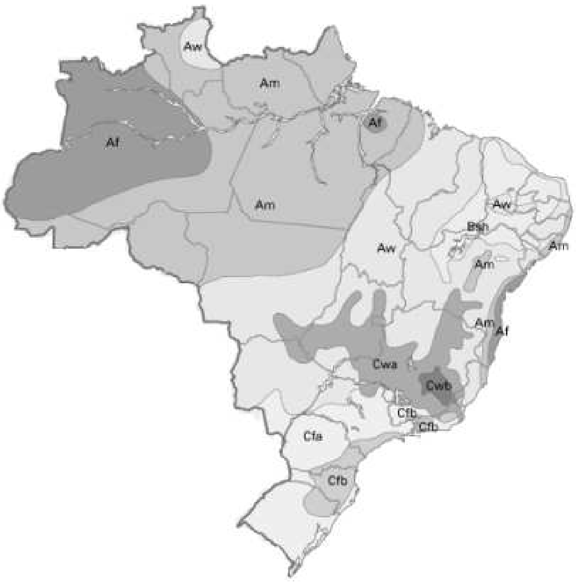

Fig. 1. Brazilian map with the Köppen climate distribution over the country.

Apache Lucene [8] search engine. The results show an enhance in precision for the same recall measures. The ontologies are considered as crisp ones where relationships between their own concepts assume values in the set $\{0, 1\}$ denoting the existence (1) or absence (0) of the relationship between them. The relationships between distinct ontologies concepts are calculated as fuzzy ones.

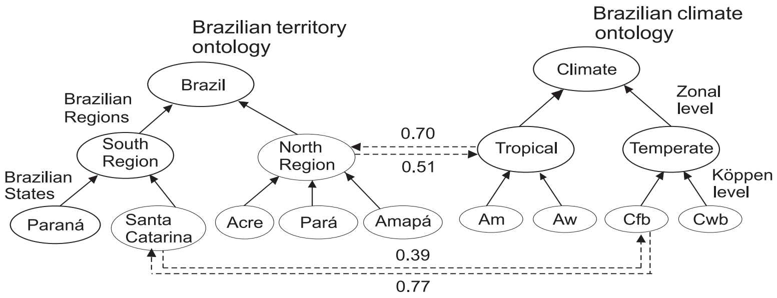

Fig. 2. Brazilian territory and climate crisp lightweight ontologies related by fuzzy associations.

2 Related Works Based on Knowledge

Some information retrieval models that encode knowledge among terms in order to improve performance are presented. In context sensitive semantic query expansion [9] a semantic encyclopedia is used as a means to provide semantics to user's query. The user queries and the index entries are represented with the use of semantic entities, defined in the encyclopedia. The query expansion process takes into account the query context, which is a fuzzy set of semantic entities. Simulation examples show that the expansion method successfully performed in direction the query context specifies. By combining lexico-syntactic and statistical learning approaches a fuzzy domain ontology mining algorithm is proposed for supporting ontology engineering [10]. The work uses one ontology as knowledge base and presents studies confirming that the use of a fuzzy domain ontology leads to significant improvement in information retrieval. By using a geographical ontology the proposed information retrieval system performs a query expansion of queries with geographical context. A query parser captures geonames and spatial relationships, and maps geographical features and feature types into concepts of a geographical ontology [11].

The multi-relationship fuzzy concept network information retrieval model [6, 7] considers the knowledge encoded as a fuzzy conceptual network. In the network each node can be related to another one by three relation types $\mathrm{V}_r : \mathrm{C}\times\mathrm{C} \rightarrow [0,1]$ where C is the concept set and $r \in \{\mathrm{P}, \mathrm{G}, \mathrm{S}\}$ denotes the fuzzy positive association (P) , fuzzy generalization association (G) and fuzzy specialization (S) association. These relations are constructed automatically based on word cooccurrence at syntactic level in the documents. The implicit relationships between concepts are inferred calculating the transitive closure for the relations resulting relations V_r^*. The documents are associated to concepts by a fuzzy relation $\mathrm{U} : \mathrm{D} \times \mathrm{C} \rightarrow [0,1]$ where D is the document set. Using the transitive closure relations V_r^* the system infers new concepts to be associated to documents resulting the expanded document descriptor relations $\mathrm{U}_r^* = \mathrm{U} \otimes \mathrm{V}_r^*$. The query q is composed with concepts from the concept network. When a query is executed the system calculates the degree of satisfaction, $\mathrm{DS}_r\left(d_i\right)$, that document $d_i \in \mathrm{D}$ satisfies the user's query q using the expanded document descriptor relations. The degree of satisfaction that a document satisfies the user's query by different fuzzy relations are aggregated to obtain the overall satisfaction for the query. The aggregation assigns a score to each document and they are presented in decreasing order to user.

The fuzzy ontological relational model [12] considers a knowledge base as a fuzzy ontology with concepts representing the categories and the keywords of a domain. When the user enters a query, composed of concepts, the system performs its expansion and may add new concepts based on the ontology knowledge. After expansion the similarity between the query and the documents is calculated by fuzzy operations. In general, works use just one conceptual structure to encode the knowledge as the ones presented is this section. The proposed model allows the knowledge to be expressed in distinct but related lightweight ontologies and offers a way to represent each ontology independently as well as their relationships. This knowledge organization is then employed in a novel method to expand queries.

3 Information Retrieval model

3.1 Knowledge, Document and Query Representations

An ontology 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. The concepts inside an ontology are organized as a taxonomy and are related by fuzzy specialization association (S) and fuzzy generalization association (G). The fuzzy generalization association is the inverse of the fuzzy specialization association. A concept is regarded as a fuzzy generalization of another concept if it consists of that concept or it includes that concept in a partitive sense. A concept is regarded as a fuzzy specialization of another concept if it is part of that concept or it is a kind of that concept. Concepts pertaining to distinct ontologies are related by fuzzy positive association (P). The fuzzy positive association denotes concepts related by a spatial, causal or similarity relation in some contexts.

Definition 1. *Consider two distinct concept domains sets* D_i *and* D_j*.*

1. *Fuzzy positive association is a fuzzy relation:* $(\mathrm{R^P}_{ij} : \mathrm{D}_i \times \mathrm{D}_j \rightarrow [0,1])$ *not symmetric, not reflexive and not transitive.*
2. *Fuzzy generalization association is a fuzzy relation:* $(\mathrm{R^G}_i : \mathrm{D}_i \times \mathrm{D}_i \rightarrow [0,1])$ *not symmetric, not reflexive and transitive.*
3. *Fuzzy specialization association is a fuzzy relation:* $(\mathrm{R^S}_i : \mathrm{D}_i \times \mathrm{D}_i \rightarrow [0,1])$ *not symmetric, not reflexive and transitive.*

The implicit relationships between concepts from the same domain are given by the transitive closure of the fuzzy generalization and fuzzy specialization associations. The transitive closure of the associations $\mathrm{R}_i^{\mathrm{G}}$ and $\mathrm{R}_i^{\mathrm{S}}$ where $1 \leq i \leq \mathrm{K}$, results the relations $\mathrm{R}^*_{\mathrm{G}i}$ and $\mathrm{R}^*_{\mathrm{S}i}$ respectively.

Definition 2. *The transitive closure* R^* *of a fuzzy relation* R *can be determined by an iterative algorithm that consists of the following steps:*

1. *Compute* $\mathrm{R}' = \mathrm{R} \cup [we_t(\mathrm{R} \circ \mathrm{R})]$ *where* $we_t \in [0,1]$*,* $t \in \{\mathrm{G,S}\}$*;*
2. *If* $\mathrm{R}' \neq \mathrm{R}$*, rename* $\mathrm{R} = \mathrm{R}'$ *and go to step 1; otherwise* $\mathrm{R}^* = \mathrm{R}'$ *and the algorithm terminates.*

The (R∘R) means the composition between two fuzzy relations. The composition between two fuzzy relations [13] $\mathrm{P} : \mathrm{X} \times \mathrm{Y}$ and $\mathrm{Q} : \mathrm{Y} \times \mathrm{Z}$ is the fuzzy relation $\mathrm{R} : \mathrm{X} \times \mathrm{Z}$ as in (1).

$$\mathrm{R}(x, z) = (\mathrm{P} \circ \mathrm{Q})(x, z) = \max_{y \in \mathrm{Y}} \min\,[\mathrm{P}(x, y), \mathrm{Q}(y, z)] \ . \qquad (1)$$

The weight we_t, with empirical values $0 < we_t < 1$, penalizes the association strength between distant concepts in the ontology. As the distance between concepts increase their association values decrease. Concepts with higher strength value are considered to have stronger meaning association. In order to discard concepts associations with lower strength value a boundary b establishes the minimum value such that the corresponding association is to be considered.

The documents d_l are represented by the DOC set where $1 \leq l \leq \|\mathrm{DOC}\|$. A fuzzy relation $\mathrm{U}_j\left(d_l,\, c_{jy}\right) = u_{ly} \in [0,\, 1]$, where $1 \leq l \leq \|\mathrm{DOC}\|$, $1 \leq j \leq \mathrm{K}$ and $1 \leq y \leq \|\mathrm{D}_j\|$ indicates the association degree between the concept $c_{jy} \in \mathrm{D}_j$ and the document $d_l \in \mathrm{DOC}$. The relations U_j, $1 \leq j \leq \mathrm{K}$ are represented as matrices $p \times m$ where $p = \|\mathrm{DOC}\|$ and $m = \|\mathrm{D}_j\|$. The U_j fuzzy relation indicates the relevance of the concept to represent the document content. Its values are calculated following a *tf-idf* schema [1].

A query is expressed with concepts from distinct domains connected by logical operators. The query is transformed into the conjunctive normal form and is represented by sub-queries connected by the AND logical operator. Each subquery is composed by a set of concepts connected by the OR logical operator. Given the domains $\mathrm{D}_1 = \{c_{11}, c_{12}, c_{13}\}$ and $\mathrm{D}_2 = \{c_{21}, c_{22}, c_{23}, c_{24}\}$ a valid query in this format would be $q = (c_{11} \vee c_{22}) \wedge (c_{13} \vee c_{24})$. Once the query is in the conjunctive normal form each sub-query is performed independently and retrieves a document set. The intersection of the document sets is the final result of the query. Therefore, in the sequence, only aspects related to the sub-query are presented. The documents are associated to the domain concepts using distinct relations. To consider this the sub-queries are partitioned to take the concepts from each domain separately. Each partition is a set with dimension equal to the associated domain concepts number and is composed by values that indicates the presence (1) or absence (0) of the concept in the query. A sub-query q is partitioned in q_i sets where $1 \leq i \leq \mathrm{K}$ and K is the domains number. In the previous example the sub-query $q = (c_{11} \vee c_{22})$ is partitioned as $q_1 = [1\,0\,0]$ and $q_2 = [0\,1\,0\,0]$.

3.2 Query Expansion

Query expansion is performed in two phases. In the first one each partition q_i, from the initial sub-query q, is expanded to consider the relations between the domain D_i associated to the partition and the other domains from the knowledge base. For each partition q_i new $\mathrm{K} - 1$ sets are generated each one containing concepts from the others domains D_j, $j \neq i$, $1 \leq i, j \leq \mathrm{K}$ associated to concepts present in q_i. This process generates a new expanded query denoted qent. The first expansion is translated in (2). The variable i refers to the domain of the partition q_i and the variable j refers to the remaining domains from the knowledge base.

$$\mathrm{qent} = \bigcup_{i=1}^{\mathrm{K}} \bigcup_{j=1}^{\mathrm{K}} \begin{cases} q_i & j = i \\ w_{\mathrm{P}}\left(q_i \circ \mathrm{R}_{ij}^{\mathrm{P}}\right) & j \neq i \end{cases} . \tag{2}$$

To expand the query to consider other domains the fuzzy positive association $\mathrm{R}_{ij}^{\mathrm{P}}$ between concepts from the domains D_i and D_j is used. The model allows to associate a weight $w_{\mathrm{P}} \in [0, 1]$, that defines the influence the fuzzy positive association will have in the expansion. Each expansion generates a new set with domain D_j concepts. The values in the sets denote the degree the concepts from the domain D_j are related to the concepts from the partition q_i. After expansion among domains the second phase is performed. This phase expands the sub-query qent considering the knowledge inside the ontologies. This expansion generates the final transposed expanded query $\mathrm{qexp}^{\mathrm{T}}$. Equation (3) presents the expansion. The association type is given by $r \in \{\mathrm{S}, \mathrm{G}\}$.

$$\mathrm{qexp}^{\mathrm{T}} = \bigcup_{i=1}^{\mathrm{K}} \bigcup_{j=1}^{\mathrm{K}} \max \begin{cases} \begin{cases} qent_{ij}^{\mathrm{T}} \\ w_r \left(\mathrm{R}_{rj}^{*} \circ qent_{ij}^{\mathrm{T}} \right) \end{cases} & j = i \\ w_r \left(\mathrm{R}_{rj}^{*} \circ qent_{ij}^{\mathrm{T}} \right) & j \neq i \end{cases} . \qquad (3)$$

Considering the knowledge inside the domains each transposed partition $qent_{ij}^{\mathrm{T}}$, $1 \leq i, j \leq \mathrm{K}$ is expanded to take into account the fuzzy generalization and fuzzy specialization associations between the concepts from their domain D_j. The model allows to associate a value $w_r \in [0, 1]$, $r \in \{\mathrm{S}, \mathrm{G}\}$ that defines a weight to the association type. This way the expansion can be adjusted to consider more one association type than the other.

3.3 Documents Relevance

The documents relevance is given by the similarity function between the documents representation and the expanded fuzzy sub-query $qexp^{\mathrm{T}}$. The similarity is calculated by the product between the relations U_j with each partition $qexp_{ij}^{\mathrm{T}}$, as in (4), resulting the retrieved documents set V.

$$\mathrm{V} = \bigcup_{i=1}^{\mathrm{K}} \bigcup_{j=1}^{\mathrm{K}} \left(\mathrm{U}_j \times qexp_{ij}^{\mathrm{T}} \right) . \qquad (4)$$

Each relation U_j associates the collection documents to the D_j domain concepts, where $1 \leq j \leq \mathrm{K}$. The set $qexp_{ij}^{\mathrm{T}}$ represents the expansion of the concepts from the partition q_i to the D_j domain where $1 \leq i$, $j \leq \mathrm{K}$. It is constituted from the D_j domain concepts and its values indicates the degree the concepts from domain D_j are associated to the concepts in partition q_i. The arithmetic product $\mathrm{U}_j \times qexp_{ij}^{\mathrm{T}}$ indicates the documents associated to the D_j domain that are related to the q_i partition. The $\bigcup$ symbol designates union and denotes the max operator. The arithmetic product adjusts the associations of the documents to the concepts (expressed in the relations U_j) by the strength of the relationships between concepts present in $qexp_{ij}^{\mathrm{T}}$. The $\mathrm{V}\,(v_l)$ set represents all the documents in the collection and its value $v_l \in [0, 1]$ indicates the degree a document d_l, $1 \leq l \leq \|\mathrm{DOC}\|$ is similar to the initial user query. Documents with $\mathrm{V}\,(v_l) > 0$ are presented to the user in decreasing order.

4 Model Evaluation

The model evaluation uses a document collection sample referring to the agrometeorology domain in Brazil, 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. Both ontologies are manually constructed.

4.1 The Ontologies Construction

To construct the ontologies the brazilian map from Fig. 1 is considered. For both ontologies the fuzzy generalization association and the fuzzy specialization association relates the spatial relationship between the entities they refer to. As ontologies are considered as crisp ones the fuzzy generalization and the fuzzy specialization relationships assume values in the set $\{0, 1\}$ denoting the existence (1) or absence (0) of the relationship between concepts. The positive relationships between distinct ontologies concepts are calculated as fuzzy ones.

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. For the brazilian territory ontology the North Region is part of Brazil country so the fuzzy specialization association value is $\mathrm{R}_1^{\mathrm{S}}\,(\text{Brazil, North Region}) = 1.0$. This means that Brazil concept is specialized by North Region concept. As the fuzzy generalization association is the inverse of the fuzzy specialization association then $\mathrm{R}_1^{\mathrm{G}}\,(\text{North Region, Brazil}) = 1.0$. This means that North Region concept is generalized by Brazil concept. Figure 2 shows a sample of the brazilian territory ontology.

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.

The relationship between ontologies is established by the distribution of climate over brazilian territory as observed in the map. The relationship is settled in two levels. The first one is between brazilian regions and zonal climates and the second one is between brazilian states and Köppen climates. The dashed lines in Fig. 2 illustrate both relationships levels. The first level of fuzzy positive relationship is between brazilian regions and zonal climates. The value of relationships is given by mapping scanning. For example, the tropical zonal climate occurs in North Region. The amount of tropical climate in Brazil is $59,811$ pixels. The amount of tropical climate in North Region is $30,616$ pixels. So the fuzzy positive association between North Region and tropical climate is given by relation value $\mathrm{R}_{12}^{\mathrm{P}}\,(\text{North Region, Tropical}) = 0.51$. This means that North Region concept implies Tropical climate concept by fuzzy positive association with strength 0.51. On the other hand the North Region extent is $43,737$ pixels so the fuzzy positive association between tropical climate and North Region is $\mathrm{R}_{21}^{\mathrm{P}}\,(\text{Tropical, North Region}) = 0.70$. This means that Tropical concept implies North region concept by fuzzy positive association with strength value 0.70.

The second level of fuzzy positive relationship is between brazilian states and Köppen climates. The total amount of Cfb Köppen climate in Brazil is

$1,781$ pixels. The amount of Cfb Köppen climate in Santa Catarina state is 693 pixels. So the fuzzy positive association between Santa Catarina State and Cfb Köppen climate is $\mathrm{R}^{\mathrm{P}}_{12}$ (Santa Catarina, Cfb) $= 0.39$. This means that Santa Catarina concept implies Cfb Köppen climate concept by fuzzy positive association with strength value equal to 0.39. On the other hand the Santa Catarina state extent is 900 pixels so the fuzzy positive association between Cfb Köppen climate and Santa Catarina state is $\mathrm{R}^{\mathrm{P}}_{21}$ (Cfb, Santa Catarina) $= 0.77$. This means that Cfb Köppen climate concept implies Santa Catarina state concept by fuzzy positive association with strength value equal to 0.77. In the information retrieval process if a user constructs a query composed with Cfb concept the query expansion process, based on the $\mathrm{R}^{\mathrm{P}}_{21}$ relation, points that Santa Catarina concept is related with Cfb concept with strength 0.77 and this concept is added to query. Considering that the document collection is about agrometeorology domain this indicates that documents indexed with Santa Catarina concept can be possibly a relevant answer, even if the documents are not indexed with the Cfb concept itself. As Cfb köppen climate covers a fraction of 0.77 from Santa Catarina state then documents related to Santa Catarina concept can contain aspects related to Cfb Köppen climate even this concept itself is not present in the documents.

The Alignment Format Level 0 [14] allows to represent fuzzy relations between lightweight ontologies concepts. It is used to represent the fuzzy positive association between concepts from distinct ontologies. An extract from the fuzzy positive association $\mathrm{R}^{\mathrm{P}}_{12}$ between the brazilian territory ontology, D_1, and the brazilian climate ontology, D_2, is presented in the following.

```
<?xml version=’1.0’ encoding=’utf-8’ standalone=’no’?>
<!DOCTYPE rdf:RDF SYSTEM "align.dtd">
<rdf:RDF
  xmlns=’http://knowledgeweb.semanticweb.org/heterogeneity/alignment’
  xmlns:rdf=’http://www.w3.org/1999/02/22-rdf-syntax-ns#’
  xmlns:xsd=’http://www.w3.org/2001/XMLSchema#’>
<Alignment>
  <xml>yes</xml>
  <level>0</level>
  <type>**</type>
  <onto1>http://localhost:8090/OntoD1</onto1>
  <onto2>http://localhost:8090/OntoD2</onto2>
  <map>
    <Cell>
     <entity1 rdf:resource=’http://localhost:8090/OntoD1#NorthRegion’/>
     <entity2 rdf:resource=’http://localhost:8090/OntoD2#Tropical’/>
     <measure rdf:datatype=’&xsd;float’>0.51</measure>
     <relation>positive12</relation>
    </Cell>
  </map>
</Alignment>
</rdf:RDF>
```

4.2 Results Analysis

The proposed model performance is compared with a similar approach, that is, the multi-relationship fuzzy concept network information retrieval model presented in Sect. 2. The experiment also tests the use of the query expansion method in the Apache Lucene text search engine. The Apache Lucene allows boosting a search concept increasing the relevance of documents indexed by the concept.

The document collection is composed of a sample of 128 documents selected from a base of 17,780 documents from the agrometeorology domain. This sample considers documents containing just one of the concepts from the ontologies as well as a combination of concepts from both ontologies. The queries set contains 35 queries considering just one concept from each ontology or two concepts from both ontologies connected with AND or OR boolean operators. For each query the relevant documents from the sample document collection are selected by a domain expert. Several experiments were ran considering many combinations of the weights we_t, $t \in \{\mathrm{S}, \mathrm{G}\}$ and w_r, $r \in \{\mathrm{S}, \mathrm{G}, \mathrm{P}\}$. After many tests all the models showed a behavior tendency concerning the precision and recall measures. Recall is the fraction of the relevant documents which are retrieved over all relevant documents in collection related to query q and precision is the fraction of the retrieved documents which are relevant to a query q over all documents in the answer set [1].

As the addition of general concepts tends to add more noise in the search results then a lower weight value is assigned to fuzzy generalization association like $w_{\mathrm{G}} = 0.3$. A higher value is assigned to fuzzy specialization association like $w_{\mathrm{S}} = 0.7$. Following the same reasoning for transitive closure calculation, the tests showed that best results are achieved with lower values assigned to weight we_{G} and higher ones to weight we_{S} like $we_{\mathrm{G}} = 0.2$ and $we_{\mathrm{S}} = 0.8$. The fuzzy positive association is tested with four different weights like $w_{\mathrm{P}} = 0.0$, $w_{\mathrm{P}} = 0.1$, $w_{\mathrm{P}} = 0.5$, and $w_{\mathrm{P}} = 1.0$.

Figure 3 presents the performance results for the models showing precision x recall curves. In these curves the best results occur when the precision values maintains high as the recall values increases. This indicates that most of relevant documents are retrieved and are presented in the top of the answer set. In the performed experiment only the best result is recorded for each model to keep the graphic understandable. The proposed model is represented by MO curves, the multi-relationship fuzzy concept network model by the CN curve and the Apache Lucene by LUC curves. In the curves legend, the KW means the use of just the entered keywords (without performing query expansion) and the numbers refer to the corresponding w_{P} value when query expansion is considered. All the models showed a behavior tendency considering the w_{P} variations.

As the proposed model and the Apache Lucene use the same query expansion method their performances have the same tendency. When considering just the user entered keywords the precision for lower recall values is high but it decreases fast as the recall values increase. When query expansion is performed the precision is high for low recall values and maintains around 45% for high recall

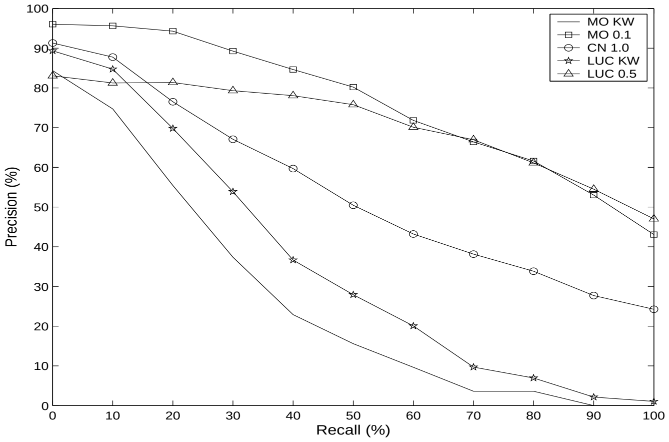

Fig. 3. Recall and precision measures for the models using crisp ontologies related by fuzzy positive association.

values. The fuzzy concept network model presents high precision values for low recall values but as the recall values are higher it maintains the precision values around 25%. Comparing the three models results, the proposed model exhibits better performance (MO 0.1 curve) when knowledge is considered in the query expansion process.

5 Conclusions

This work presents an approach for improving document retrieval process considering a knowledge base composed of multiple related domain ontologies. Contrary to other approaches that consider the knowledge base composed of just one ontology, the proposed model explores knowledge expressed in multiple ontologies that, in some contexts, can be related to each other by causal, spatial or similarity relationships. To deal with the uncertainty and vagueness present in the knowledge the fuzzy set theory is used to express the relations between concepts of distinct ontologies. This knowledge 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. When using the expansion method with the Apache Lucene search engine the results are also improved.

The knowledge organization and representation as ontologies is a growing area. Many independent developed crisp ontologies, representing distinct domains, are being proposed. The presented model offers an way where these ontologies can be reused. Instead of developing one large ontology that encodes multidisciplinary knowledge an alternative is to encode this knowledge as distinct domain ontologies relating them in a next step. This allows distinct knowledge groups to work in a independent way or to reuse already existent domain ontologies. If the ontologies represent domains that can be related in some context then just the positive fuzzy associations between the ontologies concepts have to be constructed in order to reuse them in the model. An experiment considering the ontologies themselves as fuzzy structures is presented in [15].

References

1. Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press / Addison-Wesley (1999)
2. Ogawa, Y., Morita, T., Kobayashi, K.: A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy Sets and Systems **39**(2) (1991) 163–179
3. Widyantoro, D.H., Yen, J.: A fuzzy ontology-based abstract search engine and its user studies. In: Proceedings of the IEEE International Conference on Fuzzy Systems, Washington, DC, USA, IEEE Computer Society (2001) 1291–1294
4. Bhogal, J., Macfarlane, A., Smith, P.: A review of ontology based query expansion. Information Processing and Management **43**(4) (2007) 866–886
5. Gomez-Pérez, A., Fernández-Lopez, M., Corcho, O.: Ontological Engineering. Springer-Verlag (2003)
6. Chen, S.M., Horng, Y.J., Lee, C.H.: Fuzzy information retrieval based on multi-relationship fuzzy concept networks. Fuzzy Sets and Systems **140**(1) (2003) 183–205
7. Horng, Y.J., Chen, S.M., Lee, C.H.: Automatically constructing multi-relationship fuzzy concept networks for document retrieval. Applied Artificial Intelligence, **17**(1) (2003) 303–328
8. Apache: Lucene overview http://lucene.apache.org/java/docs/index.html.
9. Akrivas, G., Wallace, M., Andreou, G., Stamou, G., Kollias, S.: Context-sensitive semantic query expansion. In: Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems (ICAIS'02), Washington, DC, USA, IEEE Computer Society (2002) 109
10. Lau, R.Y.K., Li, Y., Xu, Y.: Mining fuzzy domain ontology from textual databases. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, IEEE Computer Society (2007) 156–162
11. Martins, B., Silva, M.J., Freitas, S., Afonso, A.P.: Handling locations in search engine queries. In: GIR. (2006)
12. Pereira, R., Ricarte, I., Gomide, F.: Fuzzy relational ontological model in information search systems. In: Elie Sanchez. (Org.). Fuzzy Logic and The Semantic Web, Amsterdan, Elsevier B. V. (2006) 395–412
13. Pedrycz, W., Gomide, F.: An introduction to fuzzy sets : Analysis and Design. MIT Press, Cambridge, Massachusetts (1998)
14. Euzenat, J.: An api for ontology alignment. In: International Semantic Web Conference. (2004) 698–712
15. Leite, M.A.A., Ricarte, I.L.M.: A framework for information retrieval based on fuzzy relations and multiple ontologies. In: Iberamia 2008: Proceedings of the 11th Ibero-American Conference on Artificial Intelligence, Berlin - Heidelberg, Springer-Verlag (2008)