Semantic Query Extension using Query Contexts and Probabilistic Description Logics

José Eduardo Ochoa Luna[1], Kate Revoredo[2], and Fabio Gagliardi Cozman[1]

[1] Escola Politécnica, Universidade de São Paulo,
Av. Prof. Mello Morais 2231, São Paulo - SP, Brazil
[2] Departamento de Informática Aplicada, Unirio
Av. Pasteur, 458, Rio de Janeiro, RJ, Brazil
eduardo.ol@gmail.com,katerevoredo@uniriotec.br,fgcozman@usp.br

Abstract. A keyword-based search retrieves documents containing a set of keywords; however, documents associated to just one keyword may also be relevant. In this article we present a novel approach for semantic query extension with ontologies, using both a probabilistic description logic (PDL) and query contexts. The PDL CR$\mathcal{ALC}$ is used to model the domain associated with a collection of documents. Concepts that are related to a keyword-based query are then collected in two groups: *context concepts* associated with all keywords together and *potential context concepts* associated with keywords separately. The former group is evidence in a relational Bayesian network (RBN) built from the probabilistic description logic. Thus, documents associated with the top potential context concepts are also returned to the user as a result of the query. Examples and issues of importance in real world applications are discussed.

Keywords: Information Retrieval, Semantic Query Extension, Probabilistic Description Logic, Ontology

1 Introduction

This paper focuses on the use of ontologies, expressed through a probabilistic description logic, to improve keyword-based search. The concepts of a given ontology are taken as annotations for documents or text fragments, thus providing background knowledge and enabling intelligent search and browsing facilities. The ontological knowledge thus augments unstructured text with links to relevant concepts. For example, articles "Life of the probabilistic fish" and "A new kind of aquatic vertebrate with probabilistic processing" are all instances of the concept Publication; in a keyword-based search, the query "probabilistic fish" would return only the former paper, since the keywords "probabilistic" and "fish" appear together only in this publication. However, the keywords separately can indicate further results. An ontology can then be employed for *semantic query extension* considering the *query context*; that is, for verifying if terms associated

with the keywords individually also lead to relevant results for the query as being associated with the query context. In short, context is a set of concepts and relationships linked to a given query.

Usually there is uncertainty in such reasoning. It is often impossible to guarantee that a document is related to the query context when it is not retrieved through keyword-based search. Thus, it would be interesting if the semantic query extension system based on the query context could produce the probability of a document conditioned on the documents retrieved by the original query.

An ontology can be represented through a description logic (DL) [2], typically a decidable fragment of first-order logic that tries to reach a practical balance between expressivity and complexity. To represent uncertainty, a probabilistic DL (PDL) must be contemplated. The literature contains a number of proposals for PDLs [8, 9, 17], as this is central to the management of semantic data in large repositories. In this paper, we adopt a recently proposed PDL, called Credal $\mathcal{ALC}$ (CR$\mathcal{ALC}$) [5], that extends the popular logic $\mathcal{ALC}$[2]. In CR$\mathcal{ALC}$ one can specify sentences such as $P(\mathsf{Professor}|\mathsf{Researcher}) = 0.4$, indicating the probability that an element of the domain is a Professor given that it is a Researcher. These sentences are called *probabilistic inclusions*. Exact and approximate inference algorithms that deal with probabilistic inclusions have been proposed [5, 6], using ideas inherited from the theory of Relational Bayesian Networks (RBN) [10].

In this paper, we propose an algorithm that receives a query, performs a keyword-based search and considers semantic information about the domain of the application to obtain results that are not possible in standard information retrieval (IR). The goal is not only to retrieve documents related to the keywords, but also related to the query context, thus returning more informative results to the user needs. These documents are retrieved through the PDL CR$\mathcal{ALC}$, finding terms probabilistically related to the query context.

The remainder of this paper is organized as follows. Section 2 reviews relevant elements of IR and the PDL CR$\mathcal{ALC}$. Section 3 presents our proposed IR system. Section 4 presents some preliminary results and Section 5 concludes the paper.

2 Background

Consider, for example, that we are interested in retrieving new collaborators for working on a new paper about "Bayesian networks". A given researcher can have several near collaborators but he is likely to be unaware of all people currently working on a specific research area. Thus, he resorts to a search engine to help him find possible collaborators. A traditional search engine performs a Boolean query into documents indexed by name and abstract sections, returning documents with the keywords "Bayesian" and "network". In this section, we review the standard keyword-based information retrieval and then the PDL CR$\mathcal{ALC}$ that will be used in Section 3 to show how we retrieve other documents related to some of the keywords in the query.

2.1 Information Retrieval Models

The field of information retrieval (IR) [12] is concerned with the representation, storage, organization, and access of information items. One example of traditional IR technique is the Boolean model [16]. A document d is then represented by the vector $\overrightarrow{x} = (x_1, ..., x_M)$ where $x_t = 1$ if term t is present in document d and $x_t = 0$ otherwise. The procedure searches for documents that satisfy a query in the form of a Boolean expression of terms. Thus, if a query such as x_1 AND x_2 OR x_3 is provided, this technique retrieves documents where $x_1 = 1$ and $x_2 = 1$ simultaneously or $x_3 = 1$.

Another sort of model for IR is based on logical representations [3, 4, 11]. The task can be described as the extraction, from a given document base, of those documents d that, given a query q, make the formula $d \rightarrow q$ valid, where d and q are formulas of a chosen logic and "$\rightarrow$" denotes logical implication. In this paper we are interested in the logical representations that consider that the symbols d and q are terms (i.e. expressions denoting objects or sets of objects); accordingly, "term d is an instance of term q". Different formalisms have been proposed with these goals. An example is the terminological logic for IR proposed in [13]. In this logic, documents are represented by individual constants, whereas a class of documents is represented as a concept, and queries are described as concepts. Given a query q, the task is to find all those documents d such that $q(d)$ holds. The evaluation of $q(d)$ uses the set of assertions describing documents; that is, instead of evaluating whether d is related to q, evaluate if "individual d is an instance of the class concept q".

2.2 Probabilistic Description Logics and CR$\mathcal{ALC}$

A description logic (DL) offers a formal language where one can describe knowledge such as "A Professor is a Person who works in an Organization". To do so, a DL typically uses a decidable fragment of first-order logic [2], and tries to reach a practical balance between expressivity and complexity. The last decade has seen a significant increase in interest in DLs as a vehicle for large-scale knowledge representation, for instance in the semantic web. Indeed, the language OWL [1], proposed by the W3 consortium as the data layer of their architecture for the semantic web, is an XML encoding for quite expressive DLs.

Knowledge in a DL is expressed in terms of *individuals*, *concepts*, and *roles*. The semantics of a description is given by a *domain* Δ and an *interpretation*, that is a functor $\cdot^{\mathcal{I}}$. Individuals represent objects through names from a set of names $N_{\mathsf{I}} = \{\mathsf{a}, \mathsf{b}, \ldots\}$. Each *concept* in the set of concepts $N_{\mathsf{C}} = \{\mathsf{C}, \mathsf{D}, \ldots\}$ is interpreted as a subset of a domain $\mathcal{D}$ (a set of objects). Each *role* in the set of roles $N_{\mathsf{R}} = \{\mathsf{r}, \mathsf{s}, \ldots\}$ is interpreted as a binary relation on the domain. Objects correspond to constants, concepts to unary predicates, and roles to binary predicates in first order logic. Concepts and roles are combined to form new concepts using a set of *constructors*. Constructors in the $\mathcal{ALC}$ logic are *conjunction* ($C \sqcap D$), *disjunction* ($C \sqcup D$), *negation* ($\neg C$), *existential* restriction ($\exists r.C$), and *value* restriction ($\forall r.C$). *Concept inclusions/definitions* are denoted

respectively by $C \sqsubseteq D$ and $C \equiv D$, where C and D are concepts. Concept $(C \sqcup \neg C)$ is denoted by $\top$, and concept $(C \sqcap \neg C)$ is denoted by $\bot$.

The probabilistic description logic (PDL) CR$\mathcal{ALC}$ [6] is a probabilistic extension of the DL $\mathcal{ALC}$ that adopts an interpretation-based semantic. It keeps all constructors of $\mathcal{ALC}$, but only allows concept names in the left hand side of inclusions/definitions. Additionally, in CR$\mathcal{ALC}$ one can have probabilistic inclusions such as $P(C|D) = \alpha, P(r) = \beta$ for concepts C and D, and for role r. (If the interpretation of D is the whole domain, then we simply write $P(C) = \alpha$.) The semantics of these inclusions is roughly (a formal definition can be found in [6]) given by:

$$\forall x \in \mathcal{D} \ : \ P(C(x)|D(x)) = \alpha \qquad \text{and} \qquad \forall x \in \mathcal{D}, y \in \mathcal{D} \ : \ P(r(x,y)) = \beta.$$

We assume that every terminology is acyclic; no concept uses itself. This assumption allows one to represent any terminology $\mathcal{T}$ through a RBN, which is a directed acyclic graph. Such a graph, denoted by $\mathcal{G}(\mathcal{T})$, has each concept name and role name as a node, and if a concept C directly uses concept D, if C appear in the left and D in the right hand sides of an inclusion/definition, then D is a *parent* of C in $\mathcal{G}(\mathcal{T})$. Each existential restriction $\exists r.C$ and value restriction $\forall r.C$ is added to the graph $\mathcal{G}(\mathcal{T})$ as nodes, with an edge from r to each restriction directly using it. Each restriction node is a *deterministic* node in that its value is completely determined by its parents. Consider the following example.

Example 1. Consider a terminology $\mathcal{T}_1$ with concepts $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$. Suppose $P(\mathsf{A}) = 0.9, \mathsf{B} \sqsubseteq \mathsf{A}, \mathsf{C} \sqsubseteq \mathsf{B} \sqcup \exists r.\mathsf{D}, P(\mathsf{B}|\mathsf{A}) = 0.45, P(\mathsf{C}|\mathsf{B} \sqcup \exists r.\mathsf{D}) = 0.5$, and $P(\mathsf{D}|\forall r.\mathsf{A}) = 0.6$. The last three assessments specify beliefs about partial overlap among concepts. Suppose also $P(\mathsf{D}|\neg\forall r.\mathsf{A}) = \epsilon \approx 0$ (conveying the existence of exceptions to the inclusion of D in $\forall r.\mathsf{A}$). Figure 1 depicts $\mathcal{G}(\mathcal{T})$.

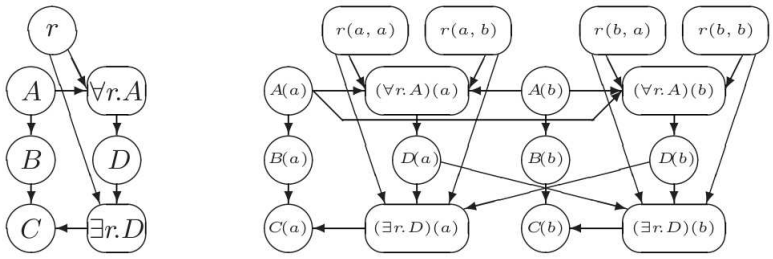

Fig. 1. $\mathcal{G}(\mathcal{T})$ for terminology $\mathcal{T}$ in Example 1 and its grounding for domain $\mathcal{D} = \{a, b\}$.

The semantics of CR$\mathcal{ALC}$ is based on probability measures over the space of interpretations, for a fixed domain. Inferences, such as $P(\mathsf{A}_o(\mathsf{a}_0)|E)$, where E is a set of observations (evidence), can be computed by propositionalization and

probabilistic inference (for exact calculations) or by a first order loopy propagation algorithm (for approximate calculations) [6]. Considering the domain $\mathsf{D} = \{a, b\}$, the grounding of $\mathcal{G}(\mathcal{T})$ of Example 1 is shown in Figure 1.

3 Semantic Query Extension using CR$\mathcal{ALC}$

We claim that a PDL such as CR$\mathcal{ALC}$ can be useful when used together with traditional IR methods. A probabilistic ontology to model the domain represented by the documents must be first created. This probabilistic ontology is represented through the PDL CR$\mathcal{ALC}$ and can be learned from data (we refer to the works on [14, 15] for detailed information on how to learn a PDL CR$\mathcal{ALC}$ from data). Then, the documents are linked to this ontology through indexes. Texts on documents are indexed and these texts are properties in the corresponding ontology. Therefore, documents and ontology are decoupled, but at the same time are related by sharing the same indexed text. The ontology and the indexed documents are input for our semantic search process. Moreover, this process receives a query that can be either (i) an *exact query* that retrieves documents that match keywords exactly, or (ii) an *open query* that retrieves documents by searching n-grams comprised in query (keywords have separate analysis). We argue for a balance between these two approaches. We assume that an exact query roughly defines a semantic context (a set of probabilistic concepts and roles linked to exact keywords denote context). Several documents retrieved by an open query could be related to the former context. Therefore, we extend an exact query by adding the most probabilistically related documents from the open queries. The semantic search process is then divided in two parts: (i) search and (ii) query extension through query context and CR$\mathcal{ALC}$. The documents are showed ranked according to their relevance. The key design choices for each task are described as follows.

Search procedure Given a query as a set of keywords, the concepts and roles related to it are found through three steps. First, an exact keyword-based search is performed finding the set of documents related to the query posed by the user. Next, the concepts and roles related to these documents are found through the corresponding indexes (therefore, the concept properties are also identified). Finally, an RBN is built where the concepts selected are evidence in this network. This RBN is the input for the query extension phase and encodes the semantic context given by the exact query.

Query extension through query context We propose to find extensions to the exact query by performing inference in the RBN obtained in the former step. Every document retrieved by the open query is probabilistically evaluated to investigate distance from the context. To do so, an individual is instantiated in the RBN representing one of the documents in the open query set (if any property is instantiated by the current document then evidence is also added accordingly). Then inference is performed for the corresponding node. The process

is repeated for each document. Then, the top related documents are retrieved. These documents are shown together with the documents related to the exact query. It is worth noting that the documents selected in the search process are re-ordered according to their probabilities, i.e., a merged ordered list of documents is exhibited to the user.

There are two main drawbacks with this proposal. The first is the size of ontologies and the second the amount of instances that are obtained after propositionalization. In principle, these issues prevent us from performing probabilistic inference on real world domains and therefore limit our framework to limited size domains. Fortunately, we can resort to variational methods in order to perform approximate inference [6].

4 Preliminary Results

Experiments were performed on a real world dataset: the Lattes Curriculum Platform[3], a public repository containing data about Brazilian researchers in HTML format. The contents of this database are quite structured (sections such as name, address education, are well defined), so that it is possible to construct a probabilistic ontology from it. We randomly selected 1964 web documents to this task, learning the probabilistic terminology from data with the CR$\mathcal{ALC}$ learning algorithm presented in [15]. The complete probabilistic terminology is given by:

$$
\begin{aligned}
& P(\mathsf{Person}) = 0.9 \\
& P(\mathsf{Publication}) = 0.5 \\
& P(\mathsf{Board}) = 0.33 \\
& P(\mathsf{Supervision}) = 0.35 \\
& P(\mathsf{hasPublication}) = 0.85 \\
& P(\mathsf{hasSupervision}) = 0.6 \\
& P(\mathsf{hasParticipation}) = 0.78 \\
& P(\mathsf{wasAdvised}) = 0.15 \\
& P(\mathsf{hasSameInstitution}) = 0.4 \\
& P(\mathsf{sharePublication}) = 0.22 \\
& P(\mathsf{sameExaminationBoard}) = 0.19 \\
\mathsf{Researcher} \equiv\ & \mathsf{Person} \\
& \sqcap (\exists \mathsf{hasPublication}.\mathsf{Publication} \\
& \sqcap \exists \mathsf{hasSupervision}.\mathsf{Supervision} \sqcap \exists \mathsf{hasParticipation}.\mathsf{Board}) \\
P(\mathsf{NearCollaborator}\ & \mid \mathsf{Researcher} \sqcap \exists \mathsf{sharePublication}.\exists \mathsf{hasSameInstitution}. \\
& \exists \mathsf{sharePublication}.\mathsf{Researcher}) = 0.95 \\
\mathsf{FacultyNearCollaborator} \equiv\ & \mathsf{NearCollaborator} \\
& \sqcap \exists \mathsf{sameExaminationBoard}.\mathsf{Researcher} \\
P(\mathsf{NullMobilityResearcher}\ & \mid \mathsf{Researcher} \sqcap \exists \mathsf{wasAdvised}. \\
& \exists \mathsf{hasSameInstitution}.\mathsf{Researcher}) = 0.98 \\
\mathsf{StrongRelatedResearcher} \equiv\ & \mathsf{Researcher} \\
& \sqcap (\exists \mathsf{sharePublication}.\mathsf{Researcher} \sqcap \\
& \exists \mathsf{wasAdvised}.\mathsf{Researcher}) \\
\mathsf{InheritedResearcher} \equiv\ & \mathsf{Researcher} \\
& \sqcap (\exists \mathsf{sameExaminationBoard}.\mathsf{Researcher} \sqcap \\
& \exists \mathsf{wasAdvised}.\mathsf{Researcher})
\end{aligned}
$$

Text on web documents was indexed according to linked properties on the ontology. When a keyword occurs within a given property, the keyword brings evidence about instance of properties for a given concept. The former probabilistic terminology acts as template for concept and property instances.

[3] http://lattes.cnpq.br/.

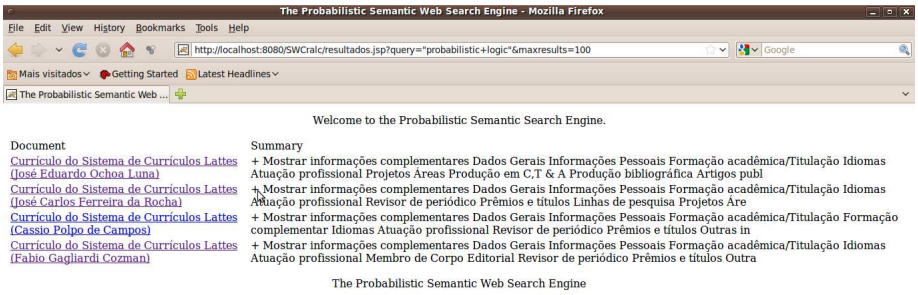

Fig. 2. Exact query results.

Assume we are interested in mapping researchers on "probabilistic logic". An exact search with a traditional search engine (Lucene[4] was used to do so) will provide, as the result to the *exact query*, a restricted set of four researchers with links to Lattes curriculum as depicted in Figure 2. On the other hand, if the keywords "probabilistic" and "logic" are analyzed separately, further results (Figure 3) are obtained, as the result to the *open query*. However, most of them can be meaningless for our goals. A probabilistic semantic contextual-awareness search balances these two extremes.

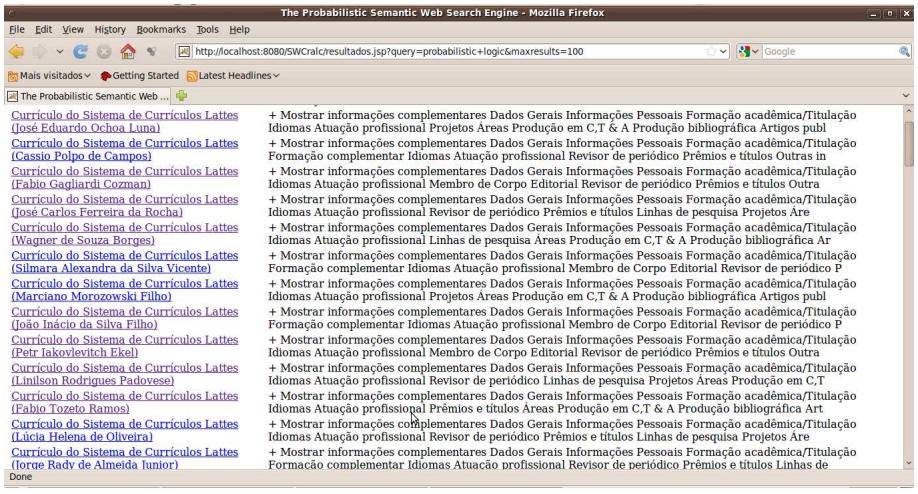

Fig. 3. Open query results.

[4] http://lucene.apache.org/

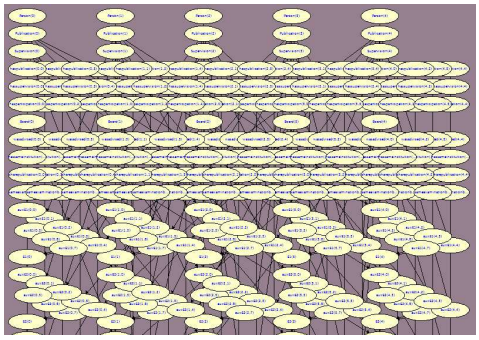

Fig. 4. Relational Bayesian network after propositionalization.

The first four results linked to the exact query bring evidence about the true context. We explore their concept and property instances in order to match context for documents retrieved from the open query so as to complete the final list of documents. We are able to instantiate specific properties, where the exact query occurs, because of indexing on text properties. This step allow us to "propositionalize" the RBN associated with the probabilistic ontology. Furthermore, in this probabilistic setting, each query occurrence inside properties denotes evidence on corresponding nodes. For instance, if a publication for $\mathsf{Researcher}(0)$ contains an exact query keyword, then the corresponding node $\mathsf{hasPublication}(0,1)$ is set to true. Some roles also allow us to state relationships among concept instances (the $\mathsf{sharePublication}(0,2)$ role relates $\mathsf{Researcher}(0)$ and $\mathsf{Researcher}(2)$ through a shared publication) and therefore enforce likelihood of related concepts that contextualize the exact query. The resulting RBN after propositionalization is shown in Figure 4.

Each document retrieved by the open query must be probabilistically evaluated to investigate proximity with the "true" context defined by the exact query. To do so, we instantiate one more individual in the RBN. This individual represents every candidate document in the open query set. We evaluate each candidate separately, i.e., evidence is set on this last indiviual (according keyword on properties) and test likelihood [5] to the context of the exact query.

In each step, just top related researchers (and their related concepts) are added to results. The reduced result page is depicted in Figure 5. Some new entries from the open query results page were added. For instance, the researcher Revoredo was correctly added because their strong related contextual information to researchers on "probabilistic logic", i.e. $(\mathsf{P}(\mathsf{Researcher}(\mathsf{Revoredo}) | \mathsf{hasPublication.P}, \mathsf{sharePublication}(\mathsf{R}, \mathsf{Revoredo})) = 0.67)$. In addition, the final research list has extended information with links to specific properties and concepts rather than uninformative snippet texts.

To evaluate results obtained by our approach we focus on searching researchers that best match several topics (given as keywords). The goal of this test

[5] Probabilistic inference is performed on the RBN.

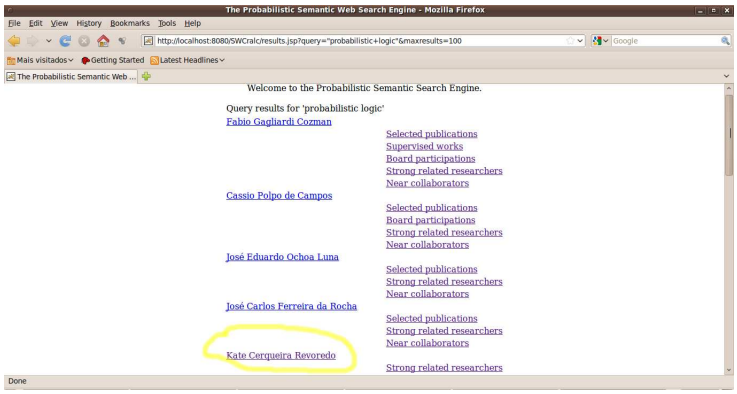

Fig. 5. Final extended result.

is to evaluate whether the semantic search return meaningful results. In order to do so, we have chosen random topics such as "Bayesian networks", "probabilistic logic", "pattern recognition" and so on with well established research groups in Brazil. Lists of researchers and related concepts were evaluated qualitatively. All 15 topics evaluated had positive analysis. Note that the analysis of results for semantic searches is still an open issue; in fact, there is no standard evaluation benchmarks that contain all required information to judge the quality of the current semantic search methods [7].

5 Conclusion

We have presented a framework for retrieving information using a mix of web documents and probabilistic ontologies. The idea is to extract semantic information in two steps. In the first step, a probabilistic ontology is constructed based on a set of documents. The second step searches for instance concepts that best match a given user query based on exact keywords — the *exact query*. The algorithm links ontology properties to indexed documents in such a way that properties are instantiated in response to queries. The resulting probabilistic ontology encodes contextual information. Further, *open query* documents are probabilistically evaluated according to their proximity to the former context. The top related are added to the final result page. Experiments focused on a real-world domain (the Lattes scientific repository) suggest that this approach does lead to improved query results.

Acknowledgements

The first author is supported by CAPES and the third is partially supported by CNPq. The work reported here has received substantial support through FAPESP grant 2008/03995-5.

References

1. G. Antoniou and F. van Harmelen. *Semantic Web Primer*. MIT Press, 2008.
2. F. Baader and W. Nutt. Basic description logics. In *Description Logic Handbook*, pages 47–100. Cambridge University Press, 2002.
3. J. Cornelis and A. van Rljsbergen. New theoretical framework for information retrieval. In *ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 194–200, 1986.
4. J. Cornelis and A. van Rljsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986.
5. F.G. Cozman and R.B. Polastro. Loopy propagation in a probabilistic description logic. In Sergio Greco and Thomas Lukasiewicz, editors, *Second International Conference on Scalable Uncertainty Management*, Lecture Notes in Artificial Intelligence (LNAI 5291), pages 120–133. Springer, 2008.
6. F.G. Cozman and R.B. Polastro. Complexity analysis and variational inference for interpretation-based probabilistic description logics. In *Conference on Uncertainty in Artificial Intelligence*, pages 1–9, 2009.
7. M. Fernandez, V. Lopez, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells. Semantic search meets the web. In *Proceedings of the 2nd IEEE International Conference on Semantic Computing*, pages 253–260, Washington, DC, USA, 2008. IEEE Computer Society.
8. J. Heinsohn. Probabilistic description logics. In *International Conf. on Uncertainty in Artificial Intelligence*, pages 311–318, 1994.
9. M. Jaeger. Probabilistic reasoning in terminological logics. In *Principals of Knowledge Representation (KR)*, pages 461–472, 1994.
10. M. Jaeger. Relational bayesian networks: a survey. *Linkoping Electronic Articles in Computer and Information Science*, 6, 2002.
11. M. Lalmas and P. Bruza. The use of logic in information retrieval modelling. *The Knowledge Engineering Review*, 13:263–295, 1998.
12. C. Manning, P. Raghavan, and H. Schütze, editors. *Introduction to Information Retrieval*. Cambridge, 2008.
13. C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–307, New York, NY, USA, 1993. ACM.
14. J. Ochoa-Luna and F.G. Cozman. An algorithm for learning with probabilistic description logics. In *5th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW) at the 8th International Semantic Web Conference (ISWC)*, pages 63–74, Chantilly, USA, 2009.
15. K. Revoredo, J. Ochoa-Luna, and F.G. Cozman. Learning terminologies in probabilistic description logics. In *Proceedings of the 20th Brazilian Symposium on Artificial Intelligence*. To appear, 2010.
16. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.
17. F. Sebastiani. A probabilistic terminological logic for modelling information retrieval. In *ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 122–130, 1994.