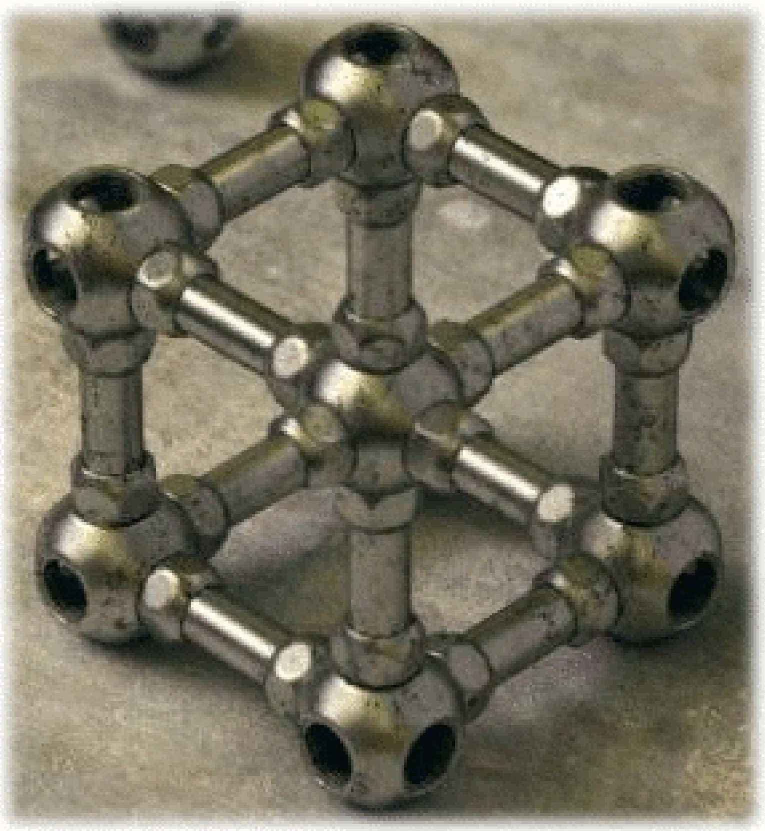


Semantic Query Extension through Probabilistic Description Logics [1]

Author(s):

José Eduardo Ochoa Luna
Kate Revoredo
Fabio Gagliardi Cozman

[1] This work was supported by Fapesp Project LogProb, grant 2008/03995-5, São Paulo, Brazil.

Semantic Query Extension through 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. This paper presents a novel approach for semantic query extension using a probabilistic description logic. Concepts that are related to a keyword-based query are used for finding other concepts and relations related through the use of a Relational Bayesian network built from the probabilistic description logic CR$\mathcal{ALC}$. Furthermore, probabilistic assessments naturally allow us to rank the information returned by search. Examples and issues of importance in real world applications are discussed.

1 Introduction

This paper focuses on the use of ontologies to improve keyword-based searches. The concepts of the given ontology are taken as annotations for documents or text fragments, thus providing background knowledge and enabling intelligent search and browsing facilities. Hence the ontological knowledge augments unstructured text with links to relevant concepts. For example, articles "Life of the probabilistic fish" and "A new kind of aquatic vertebrate with uncertainty processing" are all instances of the concept Publication; in a keyword-based search, the query "Publications about probabilistic fish" would return only the former paper. However connections amongst concepts are important to indicate further results. An ontology can then be employed for *semantic query extension*; that is, for deriving terms that lead to relevant results for the query. For example, the concept Publication is related to the concept Author; a semantic query extension strategy could use this information and reason that the second paper is a valid result as Professor G. Rouper is an author of both papers.

There is always uncertainty in this sort of reasoning. In particular, it may not be possible to guarantee that a concept is related to the ones in the query. Thus, it would be interesting if the semantic query extension system could handle the *probability* of a concept conditioned on the concepts mentioned in the query. In our example, the information about Author is valuable only if the probability of it influencing the contents of a paper is high.

An ontology can be represented through a description logic [3], which is typically a decidable fragment of first-order logic that tries to reach a practical balance between expressivity and complexity. To represent uncertainty, a probabilistic description logic must be contemplated. The literature contains a number of proposals for probabilistic description logics [10, 11, 25], as this is central to the management of semantic data in large repositories. In this paper we adopt a recently proposed probabilistic description logic, called Credal $\mathcal{ALC}$ (CR$\mathcal{ALC}$) [6], that extends the popular logic $\mathcal{ALC}$[3]. 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 [6, 7], using ideas inherited from the theory of Relational Bayesian Networks [12].

In this paper, we propose an algorithm that receives keyword-based queries and takes semantic information about the domain of the application to obtain results that are not possible in standard information retrieval. The idea here is to obtain all concept instances that are related to a given word even if that word does not appear with the concept. The system can infer relations through the probabilistic description logic CR$\mathcal{ALC}$, finding concepts probabilistically related to the ones in the query, making it possible to retrieve concepts that do not contain any of the specified words. The information related to the chosen concepts is the set of query results, and they are returned ranked by their probability.

Section 2 reviews relevant elements of information retrieval and the probabilistic description logic CR$\mathcal{ALC}$. Section 3 presents our proposal information retrieval system. Section 4 presents some preliminary experiments. Section 5 reviews some related work and Section 6 concludes the paper.

2 Background

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 the keywords in the query.

2.1 Information Retrieval Models

The field of information retrieval (IR) [14] has been defined as the subject concerned with the representation, storage, organization, and accessing of information items. One example of traditional IR technique is the Boolean model [23]. 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 [4, 5, 13]. 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 formula 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 (or: is less general than) term q". Different formalisms have been proposed with these goals. An example is the terminological logic for IR proposed in [15]. 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 [3], 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 a 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$\mathcal{ALC}$ [7] 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. For any element of the domain, the probability that this element is in C, given that it is in D is α. If the interpretation of D is the whole domain, then we simply write $P(C) = \alpha$. The semantics of these inclusions is roughly as follows (a formal definition can be found in [7]):

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

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. Considers 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 $\mathcal{G}(\mathcal{T})$.

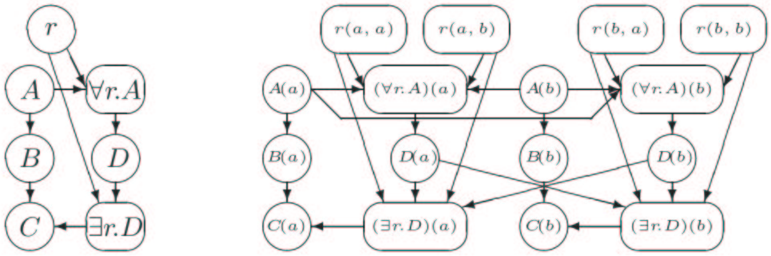


Fig. 1. $\mathcal{G}(\mathcal{T})$ for terminology $\mathcal{T}$ in Example 1 and its grounding for domain$\mathcal{D} = \{a, b\}$.

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 evidences, can be computed by propositionalization and probabilistic inference (for exact calculations) or by a first order loopy propagation algorithm (for approximate calculations) [7]. 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 through CR$\mathcal{ALC}$

In the last decade several works focusing on semantic information retrieval have been proposed. Boolean and vector space procedures, for example, have corresponding semantic versions [26, 20, 19, 8] and [27, 2, 9] respectively. We refer to [24] for a more detailed review. *Query extension* (or *query suggestion*) is a strategy often used in search engines to derive queries that are able to return more useful search results than original queries [14]. Most popular search engines provide facilities that let users complete, specify, or reformulate their queries. *Semantic* *query extension* is a special type of query extension based on the identification of semantic concepts contained in user queries [16]. For example, the result for query "Publications of probabilistic description logic" can be improved when a system that considers semantics extends the query to consider also the concept Author instead of only the concept Publication.

In [18] we used the PDL CR$\mathcal{ALC}$ combined with traditional IR to retrieve documents relevant to the query when analyzing the terms of the query separately. In this paper, we claim that the PDL CR$\mathcal{ALC}$ can also be useful for semantic query extension to obtain documents that are related to a given word even if that word does not appear with the concept. Therefore, a probabilistic ontology to model the domain represented by the documents is created. This probabilistic ontology is represented through the PDL CR$\mathcal{ALC}$ and can be learned from data (we refer to [17, 21] 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. The semantic search process is divided in three parts: (i) search, (ii) query extension and (iii) ranking the results 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, a keyword-based search is performed finding the set of documents related to the keywords provided 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, a relational Bayesian network propositionalized is built where the concepts selected are evidence in this network. This Relational Bayesian network is the input for the query extension phase.

Query Extension Procedure Expanding a given query involves adding terms and/or operators to the original query in order to improve results. In our proposal, the ontology provides terms that may be added to the query. Inference is performed in the relational Bayesian network found during search. The probability of all concepts that are not evidence in the RBN is inferred. A threshold is considered and the concepts with a probability higher than this threshold are selected and provided as input for the ranking results phase.

Ranking Procedure In this phase the documents related to the concepts selected by the query extension step are retrieved and ranked according to their probability. Then, these documents are shown together with the documents firstly selected in the search process step. It is worth noting that the documents selected in the search process are reordered according their probabilities; that is, 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 is the amount of instances that are obtained after proposizitionalization. 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 [7] making possible the application of our proposal.

4 Preliminaries Results

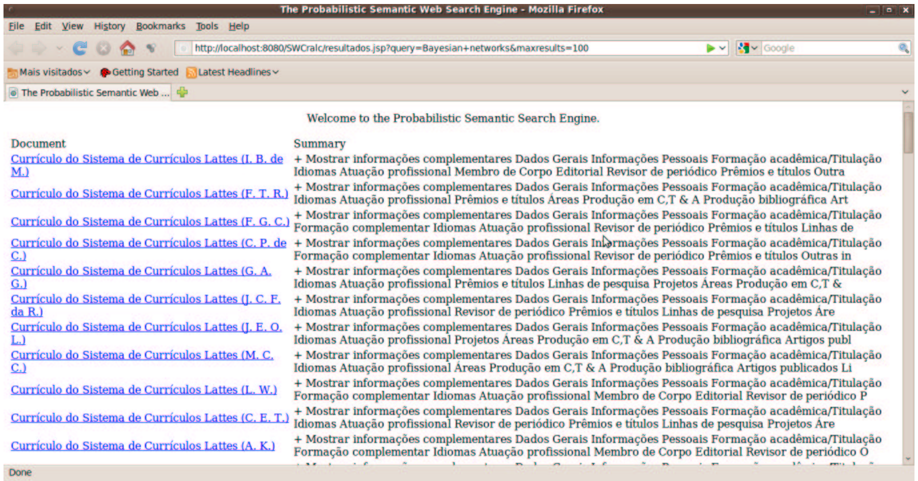
Experiments were performed on a real world dataset: the Lattes Curriculum Platform[1], a public repository containing data about Brazilian researchers in HTML format. Due its content is quite structured (sections such as name, address education, etc. are well defined) it is clearly possible to construct 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 [21]. The complete probabilistic terminology is given by:

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

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

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.

The overall process is detailed as follows. Assume we pose a query on "Bayesian networks" (the Lucene [2] search engine was used to do so), the system retrieves an ordered list of 20 researchers with links to Lattes curriculum as depicted in Figure 2.


The Probabilistic Semantic Web Search Engine - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/SWCralc/resultados.jsp?query=Bayesian+networks&maxresults=100
Google
Mais visitados
Getting Started
Latest Headlines
The Probabilistic Semantic Web ...
Done

Welcome to the Probabilistic Semantic Search Engine.

Document	Summary
Currículo do Sistema de Currículos Lattes (I. B. de M.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Idiomas Atuação profissional Membro de Corpo Editorial Revisor de periódico Prêmios e títulos Outra
Currículo do Sistema de Currículos Lattes (F. T. R.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Idiomas Atuação profissional Prêmios e títulos Áreas Produção em C,T & A Produção bibliográfica Art
Currículo do Sistema de Currículos Lattes (F. G. C.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Formação complementar Idiomas Atuação profissional Revisor de periódico Prêmios e títulos Linhas de
Currículo do Sistema de Currículos Lattes (C. P. de C.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Formação complementar Idiomas Atuação profissional Revisor de periódico Prêmios e títulos Outras in
Currículo do Sistema de Currículos Lattes (G. A. G.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Idiomas Atuação profissional Prêmios e títulos Linhas de pesquisa Projetos Áreas Produção em C,T &
Currículo do Sistema de Currículos Lattes (J. C. F. da R.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Idiomas Atuação profissional Revisor de periódico Prêmios e títulos Linhas de pesquisa Projetos Áre
Currículo do Sistema de Currículos Lattes (J. E. O. L.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Idiomas Atuação profissional Projetos Áreas Produção em C,T & A Produção bibliográfica Artigos publ
Currículo do Sistema de Currículos Lattes (M. C. C.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Idiomas Atuação profissional Áreas Produção em C,T & A Produção bibliográfica Artigos publicados Li
Currículo do Sistema de Currículos Lattes (L. W.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Formação complementar Idiomas Atuação profissional Membro de Corpo Editorial Revisor de periódico P
Currículo do Sistema de Currículos Lattes (C. E. T.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Idiomas Atuação profissional Revisor de periódico Prêmios e títulos Linhas de pesquisa Projetos Áre
Currículo do Sistema de Currículos Lattes (A. K.)	+ Mostrar informações complementares Dados Gerais Informações Pessoais Formação acadêmica/Titulação Formação complementar Idiomas Atuação profissional Membro de Corpo Editorial Revisor de periódico O

Fig. 2. Traditional query.

Suppose the user intends to follow each link and to inspect where "Bayesian networks" is located, so as to determine relevance of the document retrieved. In our setting these 20 results are candidate documents that could be further extended. Actually, these results are candidate instance concepts in the probabilistic terminology.

Furthermore, because of indexing on text properties, we are able to instantiate specific properties where the query occurs. This step allow us to "propositionalize" the inherent relational Bayesian network associated with the probabilistic ontology. Furthermore, in this probabilistic setting, each query occurrence inside properties denotes evidence on corresponding nodes. For instance, if $\mathsf{Researcher}(0)$ contains the query keyword on a given publication 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 leads to extensions of the original

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

query. The resulting relational Bayesian network after propositionalization is shown in Figure 3.

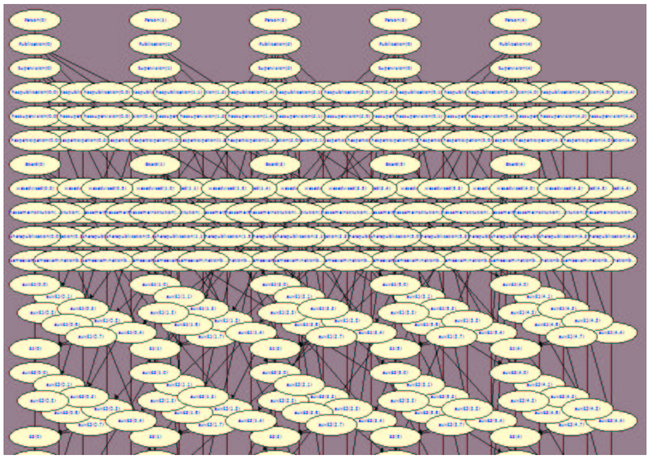
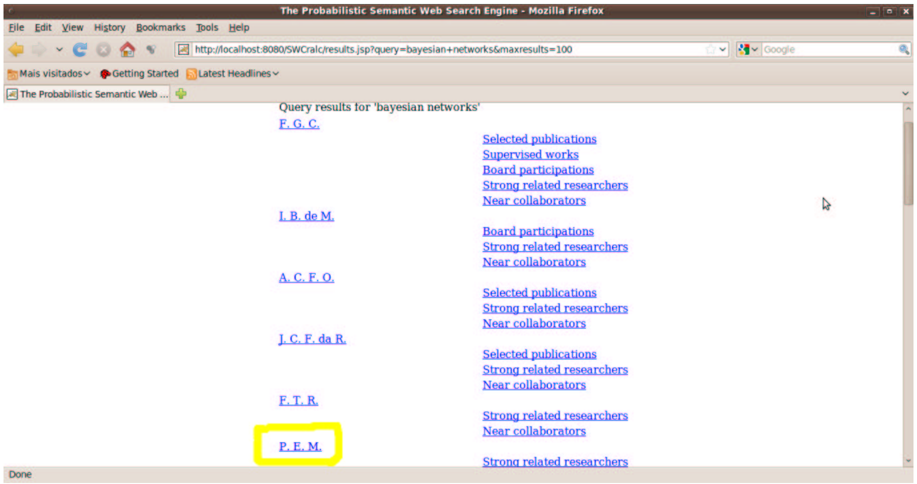


Fig. 3. Relational Bayesian network after propositionalization.

Probabilistic inference is performed on the relational Bayesian network to obtain semantic query extensions; that is, top related concepts and top related researchers to the query are added to results. The extended results page is depicted in Figure 4. Some new entries were added to the former results page (for instance, the researcher P. E. M. was added because of its strong relationship with a top researcher on "Bayesian networks"). In addition, the final research list has extended information with links to specific properties and concepts rather than uninformative snippet texts.

Probabilistic reasoning also allows us to obtain a probabilistic ranking. Intuitively, higher evidence on a given topic gives rise to a better ranking position. The previous ranking in Figure 2 returned the three following researchers: I.B. de M., F. T. R. and F.G.C. Conversely, our probabilistic logic setting returns a modified order: F.G.C., I.B. de M. and A. C. F. O. A relational Bayesian network model allow us to further investigate these results. The higher ranking was attributed to researcher F.G.C. due to evidence of query topic on publications, advising works and participations of examination boards ($\mathsf{P}(\mathsf{Researcher}(\mathsf{F.G.C.}) | \mathsf{hasPublication.P}, \mathsf{advises.S}, \mathsf{participate.B}) = \alpha$). The rest of the ranking was obtained accordingly.

To evaluate results obtained by our approach, two types of tests were conducted. The first type focuses on searching researchers that best match several topics (given as keywords). The aim of this test is 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"


The Probabilistic Semantic Web Search Engine - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:8080/SWCralc/results.jsp?query=bayesian+networks&maxresults=100
Google
Mais visitados
Getting Started
Latest Headlines
The Probabilistic Semantic Web ...
Done

Query results for 'bayesian networks'

F. G. C.

Selected publications
Supervised works
Board participations
Strong related researchers
Near collaborators

I. B. de M.

Board participations
Strong related researchers
Near collaborators

A. C. F. O.

Selected publications
Strong related researchers
Near collaborators

J. C. F. da R.

Selected publications
Strong related researchers
Near collaborators

F. T. R.

Strong related researchers
Near collaborators

P. E. M.

Strong related researchers

Fig. 4. Final extended result.

and so on with well established research groups in Brazil. Lists of researchers and related concepts were evaluated qualitatively. All 20 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 [9].

The second test addresses the ranking problem; that is, are the top researchers listed first for every topic? This issue is linked to probabilistic assessments that denote strength of relationships among instances, and give rise to a 99% positive analysis.

5 Related Work

Our framework for semantic query extension has been influenced by previous works, which we now briefly review.

The work in [22] describes a semantic search that is based on keywords, but at the same time uses the semantic information about the domain of interest to obtain results that are not possible with traditional searches. Differently from traditional searches, the work obtains all concept instances that are related to a given word even if that word does not appear inside the concept. The system can infer relations through a spread activation algorithm, making it possible to retrieve concepts that do not contain any of the specified words. The spread activation algorithm works basically as a concept explorer. Given an initial set of activated concepts and some restrictions, activation flows through the instance network reaching other concepts which are closely related to the initial concepts. One of the ideas in that work is to extract knowledge from the ontology and its instances in order to obtain a numerical weight for each existing relation instance in the model. The result is an hybrid instances network, where each relation instance has both a semantic label and numerical weight. The intuition behind this idea is that better results in the search process can be achieved using the semantic information together with the sub-symbolic (numerically encoded) information extracted from the instances. The present work is different in that it uses a relational Bayesian network to find other concepts related to the one in the query. Therefore, it also finds the probability associated to the concepts.

In [16] the most relevant concepts for the full query and for each contiguous sequence of n words of the query are collected; then, a supervised machine learning method is used to decide which of the retrieved concepts should be kept and which should be discarded. In order to train the learning algorithm, queries submitted and manually linked to relevant DBpedia concepts are used as datasets [28]. The task: given a query (within a session, for a given user), produce a ranked list of concepts from DBpedia that are mentioned or meant in the query. These concepts could then be used to suggest contextual information, such as text snippets from the Wikipedia article. One difference to the present proposal is that we do handle uncertainty explicitly; also, we do not change the original query.

Another complete framework was proposed in [9]. Basically, two tasks were addressed. The first, understanding the natural language user request and retrieving an answer in the form of pieces of ontological knowledge. The user's query is processed and translated into the terminology of available ontologies, thus retrieving a list of ontological entities as a response. In the second task, relevant documents are retrieved and ranked based on the previously retrieved pieces of ontological knowledge. Just as traditional ranking algorithms are based on keyword weighting, their approach relies on measuring the relevance of each individual association between semantic concepts and web documents. This work is related to ours because it also maintains the search process decoupled (ontology and text are explored separately). The difference relies on the consideration of uncertainty in the present work.

6 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. The algorithm links ontology properties to indexed documents in such a way that properties are instantiated in response to queries.

By handling properties and concepts we can instantiate related concepts and therefore obtain a meaningful relational Bayesian network to perform inference and to obtain a ranking of concepts. 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. K. Anyanwu, A. Maduko, and A. Sheth. Semrank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th international conference on World Wide Web*, pages 117–127, New York, NY, USA, 2005. ACM.
3. F. Baader and W. Nutt. Basic description logics. In *Description Logic Handbook*, pages 47–100. Cambridge University Press, 2002.
4. 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.
5. J. Cornelis and A. van Rljsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986.
6. 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.
7. 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.
8. L. Ding, T. Finin, A. Joshi, Y. Peng, R. Pan, and P. Reddivari. Search on the semantic web. *Computer*, 38:62–69, 2005.
9. 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.
10. J. Heinsohn. Probabilistic description logics. In *International Conf. on Uncertainty in Artificial Intelligence*, pages 311–318, 1994.
11. M. Jaeger. Probabilistic reasoning in terminological logics. In *Principals of Knowledge Representation (KR)*, pages 461–472, 1994.
12. M. Jaeger. Relational bayesian networks: a survey. *Linkoping Electronic Articles in Computer and Information Science*, 6, 2002.
13. M. Lalmas and P. Bruza. The use of logic in information retrieval modelling. *The Knowledge Engineering Review*, 13:263–295, 1998.
14. C. Manning, P. Raghavan, and H. Schütze, editors. *Introduction to Information Retrieval*. Cambridge, 2008.
15. 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.
16. E. Meij, M. Bron, B. Huurnink, L. Hollink, and M. de Rijke. Learning semantic query suggestions. In *8th International Semantic Web Conference*, pages 424–440. Springer, 2009.

17. 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.
18. J. Ochoa-Luna, K. Revoredo, and F.G. Cozman. Semantic query extension using query contexts and probabilistic description logics. In *Proceedings of the 3rd International Workshop on Web and Text Intelligence*. To appear, 2010.
19. B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. Kim – a semantic platform for information extraction and retrieval. *Nat. Lang. Eng.*, 10(3-4):375–392, 2004.
20. R. Guha R., McCool, and E. Miller. Semantic search. In *Proceedings of the 12th international conference on World Wide Web*, pages 700–709, New York, NY, USA, 2003. ACM.
21. 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.
22. C. Rocha, D. Schwabe, and M. Aragao. A hybrid approach for searching in the semantic web. In *Proceedings of the 13th international conference on World Wide Web*, pages 374–383, New York, NY, USA, 2004. ACM.
23. G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.
24. P. Scheir, V. Pammer, and S. Lindstaedt. Information retrieval on the semantic web - does it exist. In *In LWA 2007, Lernen - Wissensentdeckung - Adaptivität, 24.-26.9. 2007 in Halle/Saale (in this volume*, 2007.
25. 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.
26. A. Sheth, C. Bertram, D. Avant, B. Hammond, K. Kochut, and Y. Warke. Managing semantic content for the web. *IEEE Internet Computing*, 6(4):80–87, 2002.
27. N. Stojanovic, N. Studer, and R. Stojanovic. An approach for the ranking of query results in the semantic web. In *Proceedings of the 2nd International Semantic Web Conference*, pages 500–516, 2003.
28. I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.