KNOWLEDGE MEDIA

KMi

INSTITUTE

Document retrieval based on intelligent query formulation

Tech Report kmi-04-13

Gaston G. Burek and Maria Vargas-Vera

The Open University

Document retrieval based on intelligent query formulation

Gaston G.Burek
Knowledge Media Institute
The Open University
Milton Keynes, UK, MK7 6AA
+ 44 (0) 1908 85859
g.g.burek@open.ac.uk

Maria Vargas-Vera
Knowledge Media Institute
The Open University
Milton Keynes, UK, MK7 6AA
+44 (0) 1908 655761
m.vargas-vera@open.ac.uk

ABSTRACT
This paper presents a proposal for an open domain question answering coupled with ontological integrated space. It uses Latent Semantic Indexing (LSA) in conjunction with ontologies and First order Logic (FOL) to locate relevant documents to a query in a collection of documents. The main strength of the suggested approach relies in the use of contextual information, embedded in an integrated ontological space, to perform intelligent document retrieval.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *text analysis.* H.3.3 [**Information Storage and Retrieval]**: Information search and retrieval – *query formulation.*

General Terms
Measurement, Performance, Design, and Experimentation.

Keywords
Ontology Integration, Latent Semantic Indexing, Query Formulation, Information Retrieval, Question Answering

1. INTRODUCTION
We describe a novel methodology aiming to improve precision in automatic document retrieval. The questions in natural language are reformulated into a query containing an expanded representation of knowledge entities (i.e. ontological relations). Those knowledge entities belong to a variety of ontologies integrated in an ontological space.

Our approach involves two different knowledge representations: a) FOL predicates derived from the natural language question and b)"Pseudo" documents, temporary documents containing a description of knowledge entities.

The formulation of the query involves three steps:

- Questions formulated in English are translated to FOL using Natural Language Processing (NLP) techniques
- FOL predicates are mapped onto the ontological space by measuring their semantic similarity in relation to the knowledge entities.
- Pseudo documents (representing the knowledge entities mapped by the FOL predicates) are integrated to compose the query.

While previous studies have augmented the term-to-document matrix with additional vectors constructed from semantic structures (Guo, 2003) , our methodology stretches the capability of LSA and captures semantic similarity[1] between hierarchical information. LSA has been proven to perform better compared with the vector space model for high recall searches (Deerwester, 1990) when the vocabulary used is heterogeneous. In contrary when the vocabulary is homogeneous LSA may add noise by spurious co occurrence data producing a decrease in the precision (Manning and Schutze, 2002).

1.1 Motivation and Context
The main motivation for this work is the development of a methodology aimed to improve precision by mean of adding context information from available ontologies to a FOL predicate during query formulation process in the document retrieval phase. Although hierarchical information have been used before in query reformulation for information retrieval (Klink, 2001), such approaches replace query keywords by names of entities names that appear higher in the hierarchy of a database. Our query formulation method uses not only name of classes but also the names of properties associated with those classes.

Most of Question Answering (QA) systems are composed by four components (i.e. question analysis, document retrieval, passage retrieval, and answer extraction) (Tellex *et al.*, 2003). In particular we will concentrate in document retrieval. During the document retrieval phase documents can be retrieved by measuring semantic similarity between the query and the documents by means of using LSA and the cosine similarity measure. Ding claims that "*Dimension reduction methods, such as LSA when applied to semantic spaces built upon text collections , improve information retrieval, information filtering and word sense disambiguation"* (Ding , 2001).

Current generation of QA Systems only apply linguistic analysis techniques to the query only once the text collection is reduced to a few documents or paragraphs (Katz and Lin , 2003). This fact makes the application of query processing techniques completely redundant if the documents retrieved are not relevant to the query.

[1] Semantic similarity is measure by means of using LSA and the cosine similarity measure 0.

Low recall indicates that the level of restriction imposed to the query is too high and that restriction must be relaxed. On the other hand if the level of restriction posed on the query is too low the system precision will be also low. One way to relax the restriction posed on the query by using the cosine similarity is to expand the query adding terms that represents context knowledge. The cosine similarity measure gives the highest similarity rate to vectors that have more similar weight proportion the ones of the query. The cosine similarity measure is only determined by its topic expressed as within-object term relationship (Jones and Furnas, 1987).

Given the envision of a scenario where the Semantic Web is the main repository of knowledge we are currently researching towards the development of a methodology that combines the use of knowledge semantically structured in domain ontologies with Natural Language Processing (NLP) techniques such as Latent Semantic Analysis (LSA) for measuring semantic similarity between the query and the document collection.

In section 2 we describe an ontology integration method and how knowledge entities are represented within the ontological space , in section 3 we present our suggested architecture and methodology for intelligent query formulation, section 4 describes experimental results in mapping FOL predicates onto the integrated ontological space , the as use of LSA to create an automatic mapping between knowledge entities within different ontologies and finally in section 5 we present our conclusions and further work.

2. AN INTEGRATION METHOD TO BUILD THE ONTOLOGICAL SPACE

A collection of "pseudo" documents is created for each of the classes within the ontologies describing the domains tackled in the essay. The ontologies are described quantitatively using probabilistic knowledge (Florescu *et al*., 1997).

Each of these documents contains information (name, properties and relations) about a class. The documents are represented by a vector space model (Salton *et al.*, 1971) where each column in the term-to-document matrix represents the ontological classes and the rows represent terms occurring in the pseudo documents describing those knowledge entities.

Relations within the available ontologies are represented also by a vector space model where the columns in the term–to–document matrix are a combination of two or more vectors from the term–to–document matrix representing classes. Each column represents the relation held between the combined classes. A new column representing the binary relation derived from the question is added to the term-to-document matrix: this new column contains the weighted frequencies of terms appearing as arguments within the relation. For each question, one or more FOL predicates are derived through parsing. For instance: given the query "Do koalas live in the jungle?" the binary relation is <u>live in</u> (koala, jungle). In the case of this example, the vector representing the question contains a frequency of one in the rows corresponding to the terms koala and jungle.

2.1 Representation of Knowledge entities using pseudo documents

The "pseudo" documents represent knowledge entities belonging to the set of available ontologies. The documents are represented by a vector space model where each column in the matrix represents the classes and the rows represent terms occurring in the pseudo documents describing those knowledge entities. The entries in the term-to-document matrix are the frequency in which each term occurs in each document. Relations within the available ontologies can also be represented by the sum of the columns representing the relates classes.

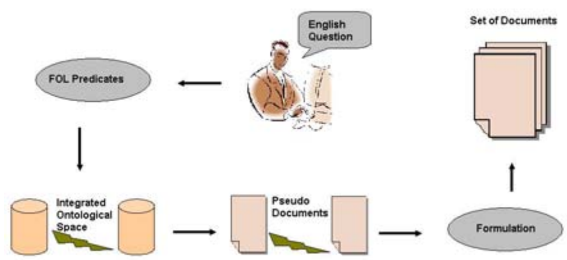

Figure 1–Achitecture for intelligent query formulation

3. ARCHICTURE FOR INTELLIGENT QUERY FORMULATION

Our Architecture for query formulation (see figure 1)[2] involves three steps: deriving the FOL predicates from the question stated in English, mapping the FOL predicates into the integrated ontological space using LSA and composing the query by means of integrating the pseudo documents. In the following subsections we will describe each of the steps in turn.

3.1 Deriving FOL predicates from the question

To derive FOL predicates from the question formulated in natural language we translate the English question into its logical form. As in AQUA[3] (Vargas-Vera *et al*., 2003) (Vargas-Vera and Motta, 2004) translation rules are used to create the logical forms.

Translation rules are used when creating the logical form of the query from grammatical components. The set of translation rules we have devised is not intended to be complete, but it does handle all the grammatical components produced by our parser. Note that variables are denoted by strings starting with a **?**, for example, **?t**.

---

[2] The notation used in the diagram is as follows: arrows represent the flow of control and ellipses represent processes.

[3] **A**utomated **Qu**estion **A**nswering System developed at Knowledge Media Institute, The Open University, UK.

The form of the logical predicates introduced by each syntax category is described as follows:

- **Nouns** (**without complement**) introduce a predicate of arity 1. For example the noun capital introduces the predicate capital (?x :type ?t_1) which restricts the type of value ?x to be the name of the city.

- **Nouns** (**with complement**) introduce a predicate of arity equal to the number of complements plus one. The pattern for n complements is as follows:

pred_name(? $argument_1$: type ?t_1, ...,? $argument_n$: type ?t_n, ? argument $_{n+1}$: type ?t $_{n+1}$).

For example, in the question ``What is the population of the UK?" the noun population is translated into the predicate:

population(uk: ?type t_1, ?x: type ?t_2).

- **Qualitative adjectives** introduce a predicate of arity 1. For example, the adjective ``AKT technology" translates into

akt_technology(?x : type ?t_1).

- **Quantitative adjectives** introduce a binary predicate. For example, the question ``How big is London?" translates into the following predicate:

has-size(london: type ?t_1, ?t :type ?t_2).

- **Prepositions** introduce a binary predicate. The pattern is as follows:

name_preposition(?$argument_1$: type ?t_1, ? $argument_2$: type t_2).

For example, the preposition *between* gets translated in the predicate:

between(?x : type t_1, ?y : type t_2).

- **Verbs** introduce predicates with one or more arguments. The first argument should be the subject of the verb, the second is the direct object, the third is the indirect object (if any) and complements (if any). For example, ``*David Brown visited KMi?*" is translated into the following predicate:

visited(david_brown: type ?t_1 , kmi: type ?t_2).

3.2 Mapping FOL predicates onto the integrated ontological space using LSA and the cosine similarity measure

In the vector space model, a term-to-document matrix is built in which the entries are weighted frequencies of pre-processed terms occurring in a collection of documents. Dimension reduction methods (such as LSA), when applied to the semantic vector space model, improve information retrieval, information filtering and word sense disambiguation. The reduction in dimensions reduces the noise in text categorisation, reduces the computational complexity of cluster creation, and produces the best statistical approximation to the original vector space model. Likelihood curves characterise with a quantity the level of significance of the reduced model dimensions. Also, the significance of each dimension follows a Zipf distribution (Li, 1992) indicating that the reduced model dimensions represent latent concepts (Ding, 2001). The dimensions in the reduced vector space model can be compared measuring semantic similarity between each of them by means of the cosine similarity. The cosine of the angle between two vectors is defined as the inner product between the vectors **v** and **w** divided by the product of the length of the two vectors.

$$Cos\theta = \frac{v.w}{\|v\|.\|w\|}$$

Given the term–to–document matrix containing a frequency f_{ij} ,the occurrence of a term in all the pseudo documents j is weighted to obtain matrix a weighted term-to-document matrix . The entries of matrix are defined as

$$a_{ij} = l_{ij} g_{ij} d_j ,$$

where l_{ij} is the local weight for term i in the pseudo document j, g_j is the global weight for term i in the collection and d_{ij} is a normalisation factor. Then, as defined by Guo (Guo , 2003),

$$a_{ij} = \log_2\left(f_{ij}+1\right)\left(1+\frac{\sum_j p_{ij}\log_2\left(p_{ij}\right)}{\log_2\left(n\right)}\right),$$

where,

$$p_{ij} = \frac{f_{ij}}{\sum_j f_{ij}} .$$

3.3 Pseudo documents integration

Once the FOL predicates have been mapped into the ontological space, the vectors representing the pseudo documents added up to conform a new vector. This vector is the final query formulation used to retrieve the subset of documents from the document collection.

Newspapers Ontology (NO)			
ID Relation	**Relation name**	**Class1**	**Class2**
OBR1	Sales Person	Advertisement	Salesperson
OBR2	Purchaser	Advertisement	Person
OBR3	Published in	Content	Newspaper
OBR4	Content	Newspaper	Content
OBR5	Employees	Organisation	Employee
OBR6	Prototype	Newspaper	Prot. Newspaper

Aktive Portal Ontology (APO)			
ID Relation	**Relation name**	**Class1**	**Class2**
OBR7	Has gender	Researcher	Gender
OBR8	Has appellation	Researcher	Appellation
OBR9	Owned by	Newspaper	Legal Agent
OBR10	Has Size	Organisation	Organisation size
OBR11	Headed by	Organisation	Afiliated Person
OBR12	Organisation part of	Organisation	Organisation

Koala Ontology (KO)			
ID Relation	**Relation Name**	**Class 1**	**Class 2**
OBR13	Has gender	Animal	Gender
OBR14	Has habitat	Animal	Appellation
OBR15	Has children	Animal	Animal

Table 1 – Ontological Binary Relations (OBR) used in Experiment

ID FOL Predicate	Argument 1	Argument 2
BP1	Advertisement	Salesperson
BP2	Advertisement	Person
BP3	Content	Newspaper
BP4	Newspaper	Content
BP5	Organisation	Employee
BP6	Newspaper	Prot. Newspaper
BP7	Researcher	Gender
BP8	Researcher	Appellation
BP9	Newspaper	Legal Agent
BP10	Organisation	Organisation size
BP11	Organisation	Afiliated Person
BP12	Organisation	Organisation
BP13	Animal	Gender
BP14	Animal	Appellation
BP15	Animal	Animal

Table 2 – Binary Predicates (BP) used in Experiment

4. EXPERIMENTAL RESULTS

The aim of this experiment is to evaluate how well LSA and the cosine similarity measure detect semantic similarity between FOL predicates and binary ontological relations integrated in the ontological space. The experiment applies the methodology described in Section 3.2 mapping the given FOL predicates onto an ontological space conformed by fifteen binary ontological relations. Those relations have been selected arbitrarily from the three available ontologies (see Table 1). The pseudo documents describing the binary ontological relations are represented as weighted terms frequencies vectors in a term-to-document matrix together with the column representing one of the Binary Predicates.

The cosine similarity (see Table 3)between binary predicates and the relations within the ontological space show that in eight cases the similarity value is higher for the relations held between classes that represent the same entities that the ones represented by the predicate arguments.

In the rest of the cases the similarity values are very close for two or more relations including the one held between classes that are that the same as the predicate arguments. Other interesting observation is that in the case of the Binary Predicate 3 (BP3) has a cosine value more similar Ontological Binary Relation 9 (OBR9), OBR3 and OBR4. In the case of the Predicate5 the cosine value is more similar to the one of OBR11 and OBR12 than for example the cosine value for OBR3 and OBR4. Similar results were obtained for the BP6 where, apart from OBR6, OBR9 has the cosine value close to one. Other similar results are repeated for BP11 and BP12 where OBR5 is more to a value of one than OBR7, OBR8 and OBR9.

	BP1	BP2	BP3	BP4	BP5	BP6	BP7	BP8	BP9	BP10	BP11	BP12	BP13	BP14	BP15
OBR1	0.3520	0.3033	0.1993	0.1993	0.2588	0.1713	-0.0007	0.0000	0.0487	-0.0030	0.0051	-0.0048	0.0000	0.0000	0.0000
OBR2	0.3628	0.3286	0.2170	0.2170	0.1896	0.1864	0.0006	0.0000	0.0528	-0.0033	0.0053	-0.0053	0.0000	0.0000	0.0000
OBR3	0.0900	0.0023	0.2864	0.2864	0.0002	0.2631	-0.0005	0.0000	0.0771	0.0017	0.0223	0.0027	0.0000	0.0000	0.0000
OBR4	0.0900	0.0023	0.2864	0.2864	0.0002	0.2631	-0.0005	0.0000	0.0771	0.0017	0.0223	0.0027	0.0000	0.0000	0.0000
OBR5	-0.0007	0.0039	-0.0013	-0.0013	0.3925	0.0000	0.0001	0.0000	-0.0006	0.0304	0.0566	0.0468	0.0000	0.0000	0.0000
OBR6	-0.0003	0.0024	0.2730	0.2730	0.0003	0.3284	0.0010	0.0000	0.0880	0.0011	0.0184	0.0017	0.0000	0.0000	0.0000
OBR7	0.0000	0.0032	-0.0004	-0.0004	0.0001	-0.0004	0.9572	0.3621	-0.0013	-0.0016	0.0143	-0.0012	0.0130	0.0130	0.0000
OBR8	0.0000	0.0032	-0.0004	-0.0004	0.0001	-0.0004	0.9567	0.3666	-0.0014	-0.0016	0.0143	-0.0012	0.0109	0.0109	0.0000
OBR9	0.0002	0.0002	0.2971	0.2971	-0.0029	0.2738	0.1477	-0.0028	0.9300	0.0147	0.0264	0.0082	-0.0002	-0.0002	0.0000
OBR10	-0.0002	0.0115	0.0014	0.0014	0.0633	0.0014	0.0999	0.0458	0.3012	0.4894	0.5181	0.3599	0.0190	0.0190	0.0000
OBR11	-0.0002	0.0113	0.0012	0.0012	0.0545	0.0012	0.1153	0.0454	0.2882	0.4304	0.4759	0.3161	0.0196	0.0196	0.0000
OBR12	-0.0002	0.0119	0.0015	0.0015	0.0522	0.0014	0.1019	0.0478	0.3146	0.4189	0.4623	0.3061	0.0221	0.0221	0.0000
OBR13	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7312	0.0000	0.0000	0.0000	0.0000	0.0000	0.5397	0.5397	0.4910
OBR14	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7490	0.0000	0.0000	0.0000	0.0000	0.0000	0.5202	0.5202	0.4767
OBR15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.7550	0.0000	0.0000	0.0000	0.0001	0.0000	0.5594	0.5594	0.5261
QBR	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3–Cosine similarity between the Binary Predicates (BP) and the Ontological Binary Relations (OBR).

The results of this experiment indicate that the presented methodology is able to detect similarity between compact representations as described by the Binary Predicates and a more expanded representations as described by the pseudo documents representing the binary relations within the three available ontologies .

Based on these results we expect that using LSA together with the cosine similarity measure we will able to pick up semantic similarity between the compacted and expanded representations of the binary relation and the document collection.

5. CONCLUSION AND FUTURE WORK

The main contribution of this paper is our outline architecture for detecting documents relevant to a query. We had showed that semantic content (encoded as ontologies) can be successfully used in query formulation.

Preliminary experiments show that semantic similarity between FOL predicates ontological relations can be successfully obtained by means of using LSA and the cosine similarity.

There is clearly a lot more work needed to make this technology work well enough for large-scale deployment.Further work may include to use our approach with different collections of documents and a large set of ontologies.

6. ACKNOWLEDGEMENTS

This work was funded by the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N157764/01.

REFERENCES

Deerwester, S.C., Dumais, S.T, Landauer ,T.K., George W. Furnas G. W. and Harshman R. A. Indexing by Latent Semantic Analysis. JASIS 41(6): 391-407 (1990)

Ding, C.H.Q. A similarity-based probability model for latent semantic indexing. Proc. 22nd ACM SIGIR Conference, pages 59-65, Aug. 1999.

Florescu, D., Koller, D. and Levy, A. Using Probabilistic Information in Data Integration Proceedings of the 23rd VLDB Conference, Athens, Greece 1997.

Guo, D., and Berry, M.W. Knowledge –Enhanced Latent Semantic Indexing .Information Retrieval. 6 (2): 225-250, 2003.

Jones, W.P. and Furnas, G.W. Pictures of relevance: a geometric analysis of similarity measures .Source Journal of the American Society for Information Science archive Volume 38 , Issue 6 (November 1987).Pages: 420 - 442.

Katz, B. and Lin, J. Selectively Using Relations to Improve Precision in Question Answering. Proceedings of the EACL 2003 Workshop on Natural Language Processing for Question Answering, April 2003, Budapest, Hungary.

Klink, S. Query reformulation with collaborative concept-based expansion. Proceedings of the First International Workshop on Web Document Analysis, Seattle, WA (2001).

W. Li. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory, 38(6):1842-1845.

Manning, C. D. and Shutze, H. Foundations of Statistical natural Language Processing . MIT Press Cambridge , Massachusetts, London, England, 2002.

Salton, G., Wong, A., and Yang, C. A vector space model for automatic indexing. Communications of the ACM, 18(11):613--620, 1971.

Tellex,S., Katz, B., Lin, J. , Marton,G., and Fernandes, A. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003), July 2003, Toronto, Canada.

Vargas-Vera, M. and Motta, E. AQUA - Ontology-based Question Answering System. Third International Mexican Conference on Artificial Intelligence (MICAI-2004), Lecture Notes in Computer Science (LNCS 2972), Springer-Verlag, April 26-30, 2004. ISBN 3-540-21459-3.

Vargas-Vera,M., Motta, E. and Domingue , J .(2003). *AQUA: An Ontology-Driven Question Answering System*. AAAI Spring Symposium, New Directions in Question Answering, Stanford University, March 24-26, 2003.