IR and AI: The role of ontology

Ying Ding

Division of Mathematics & Computer Science
Free University, Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam,
the Netherlands
Email: ying@cs.vu.nl

Abstract

This paper concerns the role of ontology in Information Retrieval (IR) and Artificial Intelligence (AI). First, it discusses the relation between IR and AI in a general way. It also gives an introduction of ontology, which could bridge the gap between IR and AI in a certain sense. This paper provides some case studies on either using IR techniques (mainly co-occurrence theory) to semi-automatically generate lightweight ontology or using already existing ontology to strengthen the retrieval. Related works has been exploited and future research has been proposed.

Keywords: information retrieval, artificial intelligence, lightweight ontology, co-occurrence theory

1. INTRODUCTION

Information overload nowadays is a well-recognised problem. With huge amount of information connected to the Internet and Intranet, efficient and effective discovery of knowledge has become an imminent research issue. Although tons of various indexing systems, cataloguing systems and searching engines are easily accessible from the web, their functions to retrieve relevant information and manage knowledge are still very limited.

There already exists several well-known problems for traditional information retrieval systems, for instance, the vocabulary inconsistency between user queries and information actually provided [1] and the simple keyword-matching approach statistically flavoured in the sense of exploiting frequency data about the occurrences and co-occurrence of natural language terms ([2], [3]).

Ontology, developed in AI to facilitate knowledge sharing and reuse, could become a possible solution [4] to eliminate these negative aspects of IR from different perspectives[5]. First, it can act at the link between users and information by logically abstracting the information so as to provide the concepts and relations (explicit semantic representation of knowledge) for users to form and refine their queries consistently. Second, it retrieves relevant information based on its inference functions, which could really fulfil the term “intelligent information retrieval”.

This paper is organised as follows. Section 2 briefly reviews some opinions about IR and AI. Section 3 gives the general introduction about ontology. Section 4 discusses two case studies focusing on the role of ontology in IR and AI. Section 5 mentions some related works. Section 6 summarises this paper and provides discussion and future works.

2. IR AND AI

Both IR and AI share the same task of finding information [6]. But they approach this task through different perspectives: representation (AI) and anti-representation (IR). AI researchers model and represent knowledge in some logical forms due to their computational tractability, explanatory power, and inference function. While, IR researchers try to retrieve information independent of any explicit data structure [7].

Actually, Sparck Jones [6] gave a good summary about the relation of IR and AI in three aspects:

- *Knowledge representation*. IR's representation of entities and relations is very weak. "Concept names are not normalised, and descriptions are mere sets of independent terms without structure … Concepts and topics, term and description meanings are left implicit… The relation between terms is only association based on co-presence..." While, the representation in AI is strong. There already exist various full-fledged methods and techniques to model the knowledge. Ontology can be considered as the generic term for generalising these representation ideas.
- *Reasoning*. The weak reasoning in IR is "looking at what is in common between descriptions and preferring one item over another because more in shared (whether as different words or, via weighting, occurrences of the same word)… The probabilistic network approach, that allows for more varied forms of search statement and matching condition, does not alter the basic style of reasoning." While development in knowledge representation of AI, especially ontology provides the backbone for reasoning and also guarantees the reasoning.
- *Learning*. Loosely speaking, the relevance feedback of IR can be considered as forms of learning. This again is very weak in IR. In this part, machine learning will link the IR and AI together to improve both sides [8].

The weakly model-based, less representation-accounted and strongly statistical methods adopted by IR have demonstrated their successes for the last decade. Now they are facing the problem to handle the information overload and other problem raised from knowledge management and electronic commerce. While ontology generated in AI area can eliminate these problems based on the semantic and machine-understandable representation of knowledge. Nowadays the manually generated ontologies cannot fulfil the increasing demands of ontologies, especially from industrial side. Semi-automatically generating, mapping and evolving ontology have become one of the hot topics in AI, which some existing full-fledged techniques in IR could contribute. On another way around, IR can further adopt ontology to refine and improve its search facilities. The aim of this paper is to use some case studies to show that ontology could link IR and AI so as to solve some problems from both sides.

3. ONTOLOGY

Since the early nineties, ontologies have become one of the popular research topics investigated by several AI communities. The reason for ontologies becoming so important is that currently we lack of standards (shared knowledge) for communication semantically and machine-understandably.

Ontology is the term referring to the shared understanding of some domains of interest, which is often conceived as a set of *classes* (concepts), *relations*, *functions*, *axioms* and *instances[9]*. Guarino [10] established a comprehensive survey of the ontology definition from the highly cited relevant works in the knowledge sharing community. Loosely speaking, any organized set of object can be considered as an ontology according to the ontology definition discussed above, for instance, catalogues, indexes from IR area; entity-relationship models (ER model) from the database community; dictionaries, thesauri from computational linguistic community; object-oriented class definition from software engineering community and so on [11].

Ontology can be represented by a language. Currently available languages are logic-based (first-order logic), frame-based (frame logic), or web-based (RDF, XML, HTML). Among them, OIL (Ontology Interchange Layer, http://www.ontoknowledge.org/oil/oilhome.shtml), a language proposed by *OntoKnowledge* Project (http://www.ontoknowledge.org) and *IBROW* (http://www.swi.psy.uva.nl/projects/ibrow/home.html), fused three paradigms: frame-based modelling with semantics based on description logic and syntax based on web standards such as XML schema and RDF schema. OIL has been successfully applied into several areas, such as knowledge management, electronic commerce and so on ([4], [12], [13]).

4. CASE STUDIES

In this part, we articulate two case studies. In the first one, we use the co-occurrence theory to detect the association relations among different keywords (concepts) so as to generate a lightweight ontology. The second one focuses on how the manually-generated ontologies could help user retrieve relevant information.

4.1 Using IR Techniques to Generate Lightweight Ontologies

In this case study, we adopted co-occurrence theory to generate lightweight domain ontology in IR. The basic assumption of co-occurrence is that if two items often co-occur together within one unit and is above some pre-defined threshold, then we believe that there exists a strong relation (or similarity) between them. For more information, please see ([14], [15]).

Firstly, literature on IR has been retrieved from the document database via the DIALOG (www.dialog.com). We select 2,012 IR documents as the sample. From each of these IR documents, we have not only accepted all the keywords added by the database indexers but important keywords from titles and abstracts as well. Finally, a total of 3,227 unique keywords were collected. Three domain thesauri were used in combination in an attempt to make the keywords consistent (singular/plural), unified (synonyms), and as far as possible, unambiguous (homonyms). Finally, 240 keywords with frequency more than two were chosen as the set of concepts (classes) for the lightweight domain ontology. A co-occurrence matrix of 240*240 keywords was formed automatically. The cell of keyword X and Y stores the co-occurrence frequency of them. We recalculated the co-occurrence frequency with the Salton Index, which is one of the important indices that can screen the negative effect of keywords with high occurrence frequency, and at the same time, reflects the direct similarity of two individual words in terms of co-occurrence frequency. In other words, this is used to eliminate high frequency words that can be linked to almost every other keyword in the research sample[16]. Based on the keywords and

their associated relations identified by co-occurrence theory, the lightweight ontology was generated. We visualised part of this lightweight ontology in Figure 1.

Figure 1. Part of visualised lightweight IR domain ontology (the dotted lines represent the link between two keywords with a Salton Index that is greater than 0.2)

First, we define two general classes: *Keyword* class and *Similarity* class. All these 240 keywords were defined as the subclass of the *Keyword* class. The *Similarity* class has three attributes: *keyword1*, *keyword2* and *weight*. The associated relation of the keyword pair (keyword1, keyword2) detected by the co-occurrence matrix was defined as the subclass of the *Similarity* class. The keyword pair (keyword1, keyword2) was the value of the *Similarity* class's attribute *keyword1* and *keyword2*, respectively. The corresponding association value was defined as the value of the attribute *weight*. Furthermore we refine this lightweight ontology with the already-existing domain thesauri to enrich the subclass relations based on the Broad Term/Narrow Term relations provided by them.

We chose the keyword "PerformanceMeasures" as the example to illustrate the structure of this lightweight IR domain ontology (see Figure 2 and Figure 3). Furthermore it was represented by the ontology representation language: OIL. It can be passed to FaCT (http://www.cs.man.ac.uk/~horrocks/FaCT/) for reasoning[1]. Other application of utilising the co-occurrence theory for ontology engineering could be that the associated keywords identified by co-occurrence could become the suggestion or recommendation list for ontology engineers when they try to create domain ontologies manually via some ontology tools (such as Protégé (http://www.smi.stanford.edu/projects/protege/protege-rdf/protege-rdf.html)).

[1] FaCT (Fast Classification of Terminologies) is a Description Logic (DL) classifier that can also be used for modal logic satisfiability testing. FaCt was developed by University of Manchester in UK. The FaCT system includes two reasoners, one for the logic SHF (ALC augmented with transitive roles, functional roles and a role hierarchy) and the other for the logic SHIQ (SHF augmented with inverse roles and qualified number restrictions), both of which use sound and complete tableaux algorithms.

4.2 Using Ontology for Improving Information Retrieval

Here we want to mention one manually-generated ontology, which could be deployed to improve information retrieval. The Knowledge Annotation Initiative of the Knowledge Acquisition Community $(KA)^2$ (part of *OntoBroker* project) is a large knowledge management initiative for the knowledge-acquisition research community which aims to develop an ontology that models this research community [17]. $(KA)^2$ ontology is a domain ontology which is used to describe the content of the information source through the notions of concepts, instances, relations, functions, and axioms. Concepts in the ontology are organised in taxonomies through which inheritance mechanisms can be applied. The $(KA)^2$ ontology consists of two main parts: a general ontology containing seven related ontologies useful for describing organisations, persons, publications, etc., and a specialised ontology for describing research topics of the knowledge acquisition community and other related scientific areas (for instance, KA through machine learning, reuse, specification languages, and so on, see Figure 4).

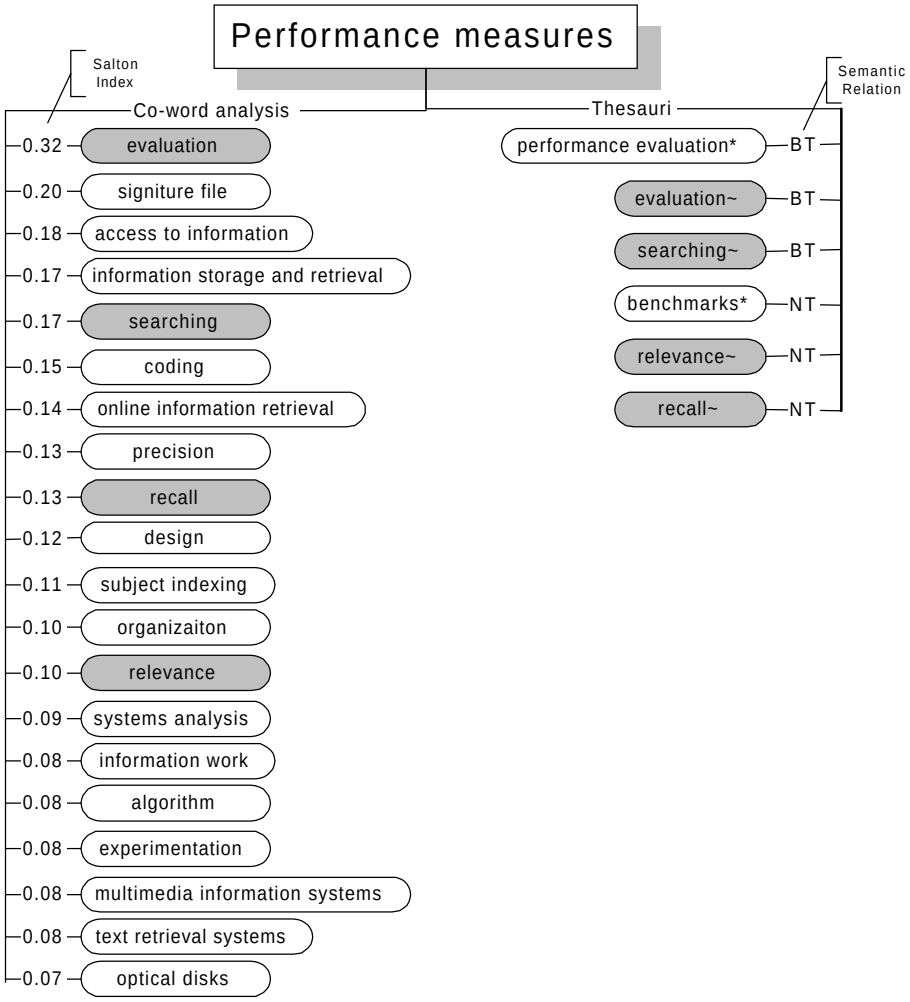

Figure 2. Part of the semantic structure of the lightweight IR domain ontology
Notes: The shaded boxes represent the similar concepts between two columns (one column represents part of the ontology generated via the co-word analysis and the other column represents part of the ontology from the three integrated existing Thesauri

```
……
class-def Keyword
class-def Similarity
  slot-constraint keyword1 value-type Keyword
  slot-constraint keyword2 value-type Keyword
  slot-constraint weight value-type Integer
……
class-def PerformanceMeasures
  subclass-of Keyword
  subclass-of Evaluation
  subclass-of Searching

class-def PerformanceMeasuresSimilarity
  subclass-of Similarity

  class-def PerformanceMeasures_Evaluation
    subclass-of PerformanceMeasuresSimilarity
    slot-constraint keyword1 has-value
PerformanceMeasures
    slot-constraint keyword2 has-value Evaluation
    slot-constraint weight has-filler 32

  class-def PerformanceMeasures_SignatureFile
    subclass-of PerformanceMeasuresSimilarity
    subclass-of Similarity
    slot-constraint keyword1 has-value
PerformanceMeasures
    slot-constraint keyword2 has-value
SignatureFile
    slot-constraint weight has-filler 20

  class-def
PerformanceMeasures_AccessToInformation
    subclass-of PerformanceMeasuresSimilarity
    slot-constraint keyword1 has-value
PerformanceMeasures
    slot-constraint keyword2 has-value
AccessToInformation
    slot-constraint weight has-filler 18
……
```

Figure 3. Part of lightweight IR domain ontology in OIL

```
class-def Organization
  subclass-of Object
  slot-constraint name  value-type STRING
  slot-constraint location  value-type STRING
  slot-constraint employs  value-type Person
  slot-constraint publishes  value-type Publication
  slot-constraint technicalReport  value-type
TechnicalReport
  slot-constraint carriesOut  value-type Project
  slot-constraint develops  value-type Product
  slot-constraint finances  value-type Project
```

Figure 4. Part of $(KA)^2$ Ontology in OIL

Figure 5 and 6 show the user interface of $(KA)^2$ to assist query formulation or refinement based on the visualised $(KA)^2$ ontology. Figure 5 is the screen shot of hyperbolic view of $(KA)^2$ ontology. Each circle represents a specific class and the subclasses of this class are linked by the solid lines. While browsing the ontology via the hyperbolic view, the selected class will automatically appear in the corresponding class slot in Figure 6. With the suggestion list of other slots (e.g. attribute slot) in Figure 6, users can easily articulate their queries or combinations of the queries through Boolean operators (AND, OR, NOT). After submitting the final query, users will get the retrieved results immediately.

Using ontology for retrieving, user could get relevant information based on the inference function of ontology. For instance, if a user wants to search a researcher's research interests or publications in this $(KA)^2$ community. Normally, he/she uses this researcher's name as the query and goes to some search engines to search. The final results would comprise lots of noisy data, for instance, the homepages of companies or other persons with the same name, some other irrelevant pages containing this name, and so on. While using this $(KA)^2$ ontology based on the defined rules or axioms, this user will get not only the right information. but other relevant information as well, for instance, researchers the searched researcher often co-operated with, the

research groups this researcher joins, the research topics he has, furthermore, the projects he participants, the funding about this project, the products produced by this project and so on [4].

5. RELATED WORKS

There are some researches having been done regarding to semi-automatically generated thesauri which White & McCain[18] considered it as the record of terms connected by co-occurrence in literature rather than lexically in language [15]. Schutze & Pedersen [14] practised co-occurrence techniques to compute thesauri from a text corpus, which significantly improve recall/precision performance over the Tipster reference corpus. Actually these IR researchers are already on their way to generate ontology (loosely speaking, all kinds of thesauri can be considered as lightweight ontologies, or linguistic ontologies (e.g. WordNet)) semi-automatically or automatically. But these researchers didn't go further for reforming these automatic thesauri to lightweight ontologies or linking them with some existing domain ontologies (if possible) to put some simple inference function in the retrieval.

Hwang [19] proposed one method for automatic generation of ontology started from the seed-words suggested by domain experts. This system collected relevant documents from the Web, extracted phrases containing seedwords, generated corresponding concept terms and located them in the 'right' place of the ontology. Several kinds of relations are extracted: is-a, part-of, manufactured-by or owned-by etc. It also collects "context lines" for each concept generated, showing how the concept was used in the text, as well as frequency and co-occurrence statistics for word association discovery and data mining. It is a nice example in a certain sense that linking IR and AI (ontology) together in order to improve the retrieval. The drawback is that it fully depends on the seedwords provided by the domain experts.

Maedche and Staab [20] proposed an approach to generate ontology semi-automatically based on the shallow text processing and learning algorithms. The outputs of the first part are dependency relations found through lexical analysis. These relations were treated as the input of the learning algorithms. Some of the dependency relations didn't hold the meaningful relations of the two concepts that could be linked together (co-occurrence) by some mediator (i.e., proposition, and so on). They also built up a system to facilitate the semi-automatic generation of the ontologies called Text-To-Onto ([20], [21]). Kietz, Maedche, and Volz [22] adopted the above method to build an insurance ontology from a corporate Intranet.

Faure and Nedellec [23] presented an interactive machine learning system called ASIUM, which is able to acquire taxonomic relations and subcategorization frames of verbs based on syntactic input. The ASIUM system hierarchically clustered nouns based on the verbs that they co-occur with and the vice versa. Byrd & Ravin [24] extracted named relations when they find particular syntactic patterns, such as an appositive phrase. They derived unnamed relations from concepts that co-occur by calculating the measure for mutual information between terms. So these researches provide some appropriate ways to extract relations among the nouns (concepts) for the target ontology.

6. DISCUSSION, CONCLUSION AND FUTURE WORK

This paper has discussed the relation between IR and AI in a general way. It also provided some case studies on either using IR techniques (mainly co-occurrence theory) to semi-automatically generate lightweight ontology or using already existing ontology to strengthen the retrieval.

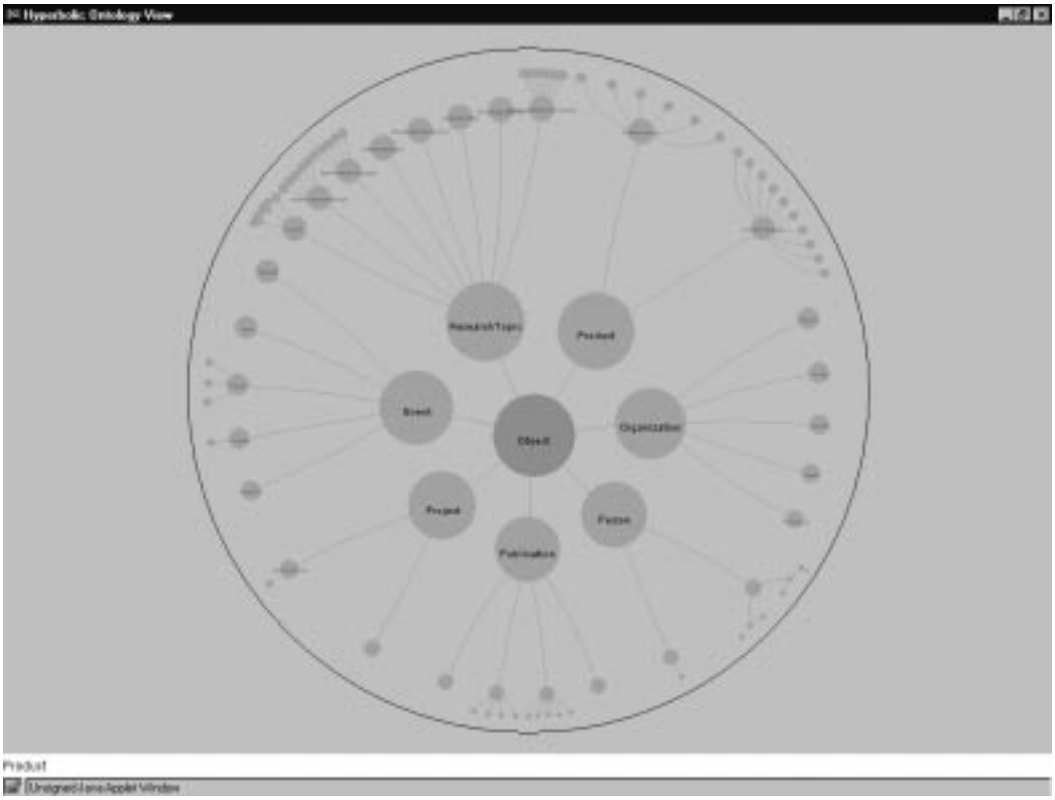

Figure 5. Hyperbolic view on (KA)^2 ontology (http://www.aifb.uni-karlsruhe.de/WBS/broker/inhalt-query.html**)**

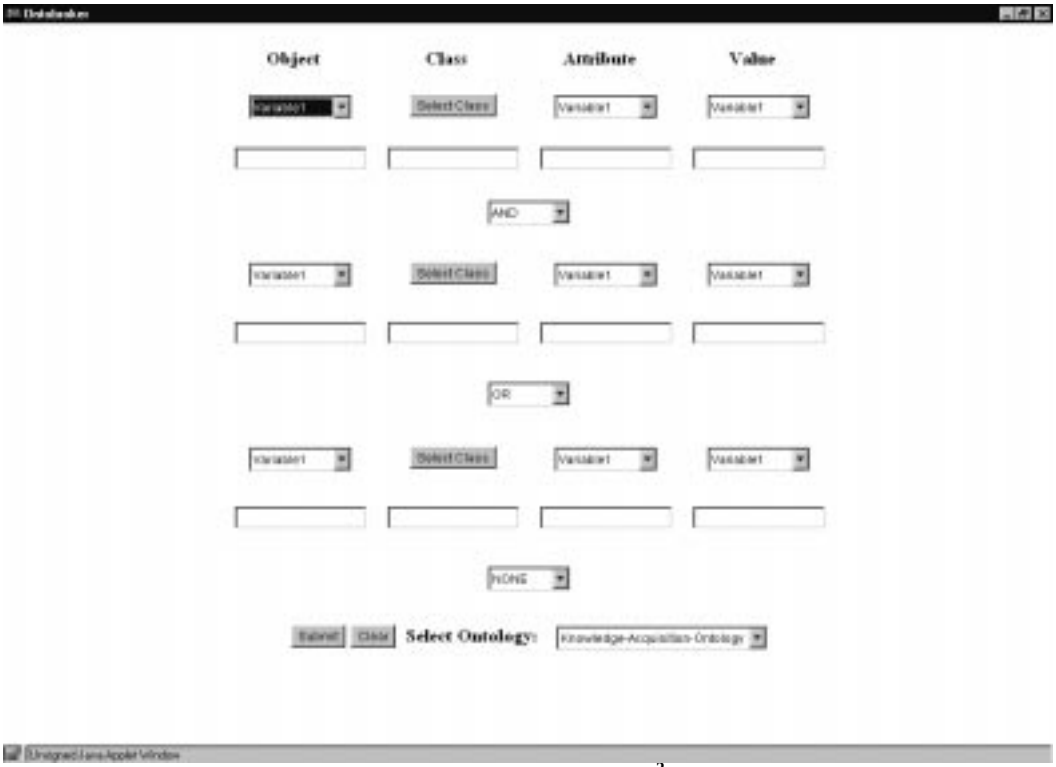

Figure 6. Query formulation interface of (KA)^2 (http://www.aifb.uni-karlsruhe.de/WBS/broker/inhalt-query.html)

Actually, AI (especially the sub-domain knowledge representation) and IR are two inseparable sides of a coin. In some aspects, they could compensate each other to eliminate negative influence. For instance, AI needs to model the knowledge explicitly and expressively through ontologies, while IR doesn't have to represent knowledge and annotate data. In addition, keyword-based IR system is fully based on the statistical matching, while AI could provide inference and subsumption to enhance the precision/recall of the whole information retrieval process.

Guarino, Masolo and Vetere [25] summarized IR and AI from the following three aspects:

- Text retrieval (partially ontology involved, no encoding for document and partially coding for query): to find the relevant document from a large collection in response to the user queries. Currently techniques based on word co-occurrence analysis integrated with morphological analysis and word stemming could generate roughly matching between documents and queries. The application of ontology here could be assistance of query formulation and refinement via browsing the ontology. Very simple or basic inference could also be provided by the ontology. Furthermore, the associated relation among keywords identified by co-occurrence analysis could become a way to semi-automatically generate lightweight ontology, for instance, the first case study.
- Data retrieval (partially lightweight ontology involved, partially encoding data and queries): both queries and data are encoded by a structured list of words (for instance, a fixed taxonomy or database schema). Data retrieval is comparatively easier than text and knowledge retrieval because either it can be easily tagged or tokenized or it is stored in strict structure.
- Knowledge retrieval (ontology involved, both data and queries are encoded): both the query and data-encoding language are much more expressive. However, the design of ontology (primitive concepts and relations) and the computational problems bound to using sophisticated knowledge representation languages might constitute a serious practical drawback. At this moment, some methods of IR could be adopted to simplify or assist the encoding process. For instance, in second case study, word co-occurrence could assist to identify the relations among concepts. Some information extraction techniques could be useful in annotating facts.

So combining IR and AI from ontology perspective will become the focus of future research ([20], [21]). Here we mention two future directions including (1) employing IR techniques for ontology learning (including ontology generation, ontology mapping and ontology evolving); (2) utilising ontologies to strengthen IR. For the first one, we want to concentrate on (1) finding proper IR or IE techniques for ontology learning based on the exploration and comparison of current IR or IE techniques; (2) conducting real-life case studies and evaluating these techniques. For the second one, efforts will be put on: (1) inference or reasoning function of ontology which could benefit current keyword-driven information retrieval system, and (2) combining current statistical IR techniques with existing relevant ontologies to improve retrieval performance.

ACKNOWLEDGEMENT

Here the author would like to thank Dieter Fensel, Michel Klein (Free University, Amsterdam) and Ian Horrocks (University of Manchester, UK) for suggestions and comments.

REFERENCE

[1] Bates, M. J. Indexing and access for digital libraries and the Internet: Human, database, and domain factors. *Journal of the American Society for Information Science*, 49(13, 1998), 1185-1205.

[2] Sparck-Jones, K., & Willett, P. *Readings in information retrieval*. Morgan Kaufmann, 1997.

[3] Ding, Y., Chowdhury, G.G., Foo, S. Incorporating the results of co-word analyses to increase search variety for information retrieval. *Journal of Information Science*, 26(6, 2000), 429-452.

[4] Fensel, D. *Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce*. Springer-Verlag, Berlin, 2001.

[5] van Harmelen, F. & Fensel, D. Practical Knowledge Representation for the Web. In. Fensel, D. (Ed.). *Proceedings of the IJCAI'99 Workshop on Intelligent Information Integration*, 1999.

[6] Sparck Jones, K. Information retrieval and artificial intelligence. *Artificial Intelligence*, 114 (1999), 257-281.

[7] Wilks, Y. IR and AI: traditions of representation and anti-representation in information processing, 2000.

[8] Michalski, R. & Kaufmann, K. Data mining and knowledge discovery: A review of issues and multi-strategy approach. In., *Machine Learning and Data Mining Methods and Applications*. John Wiley, England., 1998.

[9] Gruber, T.R.. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(1993), 199-220.

[10] Guarino, N. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In M. T. Pazienza (ed.), *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology* (pp.139-170). Springer Verlag., 1997.

[11] Wiederhold, G. & Jannink, J. Composing diverse ontologies. In. *IFIP Working Group on Database 8th Working Conference on Database Semantics (DS-8)*, in Rotorua, New Zealand, DS-8, 1999.

[12] Fensel D., Horrocks, I., Van Harmelen, F., Decker, S., Erdmann, M. & Klein, M. OIL in a nutshell. In *Knowledge Acquisition, Modeling, and Management, Proceedings of the European Knowledge Acquisition Conference (EKAW-2000)*, In. Dieng, R. et al. (eds.), *Lecture Notes in Artificial Intelligence*, LNAI, Springer-Verlag, October 2000.

[13] Klein, M., Fensel, D., van Harmelen, F. & Horrocks, I. The relation between ontologies and schema-languages: Translating OIL-specifications in XML-Schema. In *Proceedings of the Workshop on Applications of Ontologies and Problem-solving Methods, 14th European Conference on Artificial Intelligence ECAI'00*, Berlin, Germany August 20-25, 2000.

[14] Schutze, H., & Pedersen, J. O. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33(3, 1997), 307-318.

[15] Chen, H., Martinez, J., Kirchhoff, A., Ng, T. D., & Schatz, B. R. Alleviating search uncertainty through concept associations: Automatic indexing, co-occurrence analysis, and parallel computing. *Journal of the American Society for Information Science*, 49(3, 1998), 206-216.

[16] Noyons, E.C. M. & van Raan, A.F. J. Monitoring scientific developments from a dynamic perspective: self-organized structuring to map neural network research. *Journal of the American Society for Information Science*, 49(1, 1998), 68-81.

[17] Benjamins, V. R.; Fensel, D.; Decker, S. and Gomez Perez, A. $(KA)^2$: Building Ontologies for the Internet: a Mid Term Report, *International Journal of Human-Computer Studies (IJHCS)*, 51(1999):687-712.

[18] White, H. D., & McCain, K. W. Visualisation of literatures. *Annual Review of Information Science and Technology*, 32(1997), 99-168.

[19] Hwang, C. H. Incompletely and imprecisely speaking: Using dynamic ontologies for representing and retrieving information. In. *Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99)*, Linköping, Sweden, July 29-30, 1999.

[20] Maedche, A. & Staab, S. Mining Ontologies from Text. In: Dieng, R. & Corby, O. (Eds). *EKAW-2000 - 12th International Conference on Knowledge Engineering and Knowledge Management*. October 2-6, 2000, Juan-les-Pins, France. LNAI, Springer.

[21] Maedche, A. & Staab, S. Semi-automatic engineering of ontologies from text. In. *Proceedings of the 12th internal conference on software and knowledge engineering.* Chicago, USA, KSI, 2000.

[22] Kietz, J.-U., Maedche, A. and Volz, R. (Extracting a Domain-Specific Ontology Learning from a Corporate Intranet. *Second "Learning Language In Logic" LLL Workshop, co-located with the International Conference in Grammere Inference (ICGI'2000) and Conference on Natural Language Learning (CoNLL'2000)*. Lisbon, Portugal, September 13-14, 2000.

[23] Faure, D. & Nedellec, C. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In. *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*. Granada, Spain, 1998.

[24] Byrd, R. & Ravin, Y. Identifying and extracting relations from text. In. *NLDB'99 – 4th International conference on applications of natural language to information systems*, 1999.

[25] Guarino, N., Masolo, C., & Vetere, G. OntoSeek: Content-based access to the Web. *IEEE Intelligent Systems*, (May/June, 1999), 70-80.