IR and AI: Using Co-occurrence Theory to Generate Lightweight Ontologies

Ying Ding
Depart of Mathematics & Computer Science,
Vrije Universiteit Amsterdam, Nederland (ying@cs.vu.nl)

Rob Engels
CognIT, Norway (rob@cognit.no)

Abstract

This paper illustrated the application of co-occurrence theory to generate lightweight ontologies semi-automatically. First, the relationship of Information Retrieval (IR) and Artificial Intelligence (AI) is discussed in a general way. Then two case studies have been conducted to generate lightweight ontologies in specific domains (Information Retrieval domain and European part of CIA FactBook). Further discussion was articulated and future work was proposed, especially the possible future research direction on ontology learning.

1. Introduction

The emergence of the Internet has brewed the revolution of information storage and retrieval. Although various indexing, cataloguing systems and searching engines are easily accessible from the web, their functions to retrieve relevant information and manage knowledge are still very limited. Therefore, effective or intelligent search for information on this massive information resource becomes highly critical [1].

Information retrieval (IR) is concerned with the processes involved in the presentation, storage, searching and finding of information relevant to user's information needs [2]. While artificial intelligence (AI) is the sub-field of computer science concerned with the concepts and methods of symbolic inference by computer and symbolic knowledge representation for use in making inferences [3]. In IR, representing information aims at re-organising or re-ordering heterogeneous data so as to speed up the accessibility and shorten the retrieval time. So there is no requirement for the representation format to be very strict (logical). That is why normally natural language (with simple stemming or statistical association) or controlled vocabulary (normalised natural language) is often chosen for indexing and cataloguing in IR. While in AI, representation of knowledge aims at the machine semantically understanding of information in the sense that computer can generate self-learning or reasoning as well [4]. This is also the target of the "semantic web"[5]. So to represent knowledge logically and semantically is the critical requirement for the Semantic Web.

The crux is whether IR and AI can help each other to achieve the final target (finding information) effectively and efficiently. This paper implemented this crux on how to utilise co-occurrence theory of IR to generate lightweight ontology semi-automatically. It contains the following sections. Section 2 generally introduces the co-occurrence theory of IR. Section 3 discusses some case studies concerning real examples of utilisation of co-occurrence for the generation of lightweight ontologies. Section 4 mentions some relevant works. Section 5 closes the paper with conclusion and future works.

2. IR: co-occurrence

The basic assumption of co-occurrence in IR is that if two items often co-occur together within one set of document then the strong similarity exists between them. Early experiments demonstrated the potential of co-occurrence data for the identification of search term variants. Co-occurrence is normally used to expand query by means of thesauri or associative thesauri based on probabilistic models [6]. A series of recently studies have successfully employed co-occurrence to generate domain-specific thesauri semi-automatically [7].

3. Case studies: lightweight ontology

In this section, two concrete case studies are illustrated so as to show how the co-occurrence theory can be deployed to generate specific lightweight ontologies semi-automatically.

3.1 Information Retrieval

In this case study, we adopted co-occurrence theory to generate lightweight domain ontology in IR. 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. 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 [8].

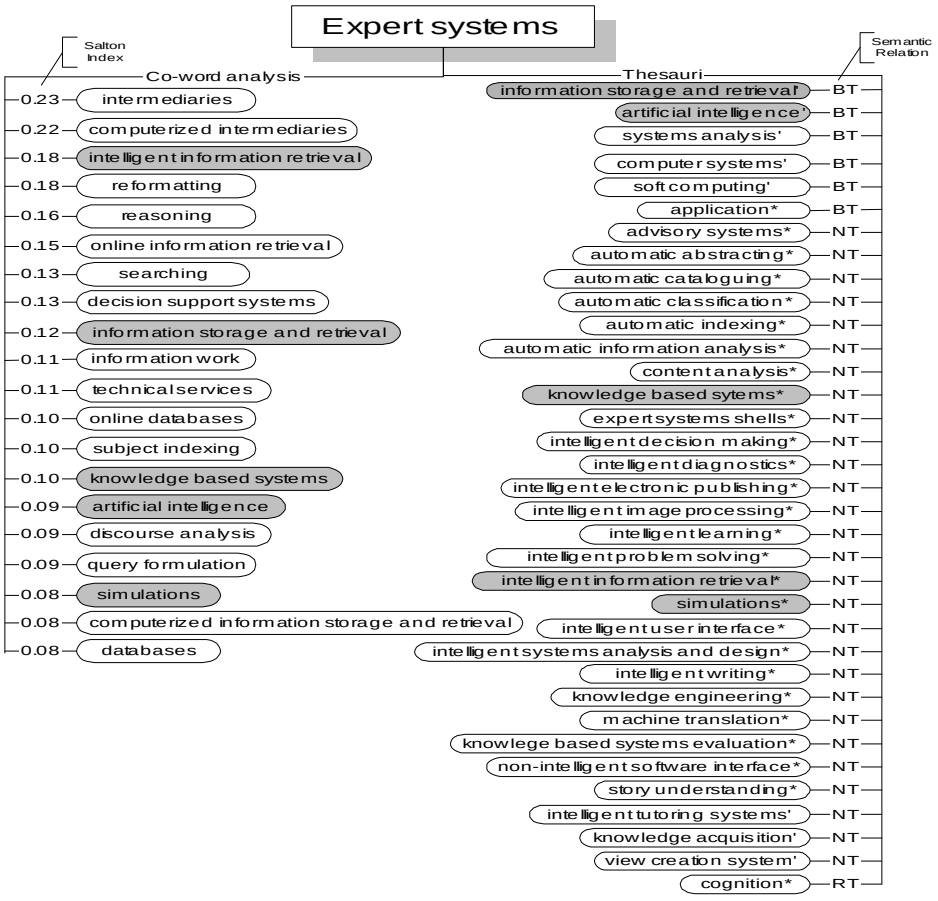

Figure 1. Part of the semantic structure of the lightweight IR domain ontology

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 "ExpertSystem" as the example to illustrate the structure of this lightweight IR domain ontology (see Figure 1). Furthermore this lightweight IR domain ontology was defined by using

OIL (Ontology Inference Layer, http://www.ontoknowledge.org/oil/oilhome.shtml) [9] (see Figure 2). It can be passed to FaCT (http://www.cs.man.ac.uk/~horrocks/FaCT/) for reasoning.

```
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 ExpertSystems
  subclass-of Keyword
  subclass-of InformationStorageAndRetrieval
  subclass-of ArtificialIntelligence

  class-def ExpertSystemsSimilarity
    subclass-of Similarity

    class-def ExpertSystems_Intermediaries
      subclass-of ExpertSystemsSimilarity
      slot-constraint keyword1 has-value ExpertSystems
      slot-constraint keyword2 has-value Intermediaries
      slot-constraint weight has-filler 23

    class-def ExpertSystems_ComputerizedIntermediaries
      subclass-of ExpertSystemsSimilarity
      slot-constraint keyword1 has-value ExpertSystems
      slot-constraint keyword2 has-value
ComputerizedIntermediaries
      slot-constraint weight has-filler 22

    class-def ExpertSystems_IntelligentInformationRetrieval
      subclass-of ExpertSystemsSimilarity
      slot-constraint keyword1 has-value ExpertSystems
      slot-constraint keyword2 has-value
IntelligentInformationRetrieval
      slot-constraint weight has-filler 18
```

Figure 2. Part of lightweight IR domain ontology in OIL

3.2 CIA FactBook

Currently, there are various kinds of software tools supporting text mining. The common features of these tools are the template-mining, information summary or information filtering based on information retrieval, information extraction, or natural language processing techniques.

Here we want to mention several software tools (sponsored by IST *On-to-Knowledge* Project, www.ontoknowledge.org) as examples. Corporum is a tool designed by CognIT (Norway) that is able to extract content representation models from natural language texts and use these models for information summary and information retrieval. Corporum can identify important concepts (keyword, or noun phrases) from heterogeneous or semi-structure information sources. Currently, they conducted several case studies to generate lightweight specific ontologies by using Corporum and represented these ontologies in XML or other ontology languages in the foreseeable future, for instance, OIL (Ontology Inference Layer), RDF and so on.

```
class-def Concept
class-def Relation
  slot-constraint concept1 value-type concept
  slot-constraint concept2 value-type concept
  slot-constraint strength value-type Integer
class-def Transportation
  subclass-of Concept
class-def RailTransportationEquipment
  subclass-of Concept
class-def Waterway
  subclass-of Concept
class-def CommonCarrier
  subclass-of Concept
class-def TransportationRelation
  subclass-of Relation

  class-def Transportation_RailTransportationEquipment
    subclass-of TransportationRelation
    slot-constraint concept1 has-value Transportation
    slot-constraint concept2 has-value
RailTransportationEquipment
    slot-constraint strength has-filler 70
  class-def Transportation_Waterway
    subclass-of TransportationRelation
    slot-constraint concept1 has-value Transportation
    slot-constraint concept2 has-value Waterway
   slot-constraint strength has-filler 40
  class-def Transportation_CommonCarrier
    subclass-of TransportationRelation
    slot-constraint concept1 has-value Transportation
    slot-constraint concept2 has-value CommonCarrier
    slot-constraint strength has-filler 40
```

Figure 3. Part of lightweight ontology about European countries in OIL

WebMaster is a software tool designed by AIdministrator (Netherlands) with the capability of analysis the contents of weakly or semi-structured information sources and visualisation of the final results. The visualisation function of WebMaster can denote the hidden cluster relation among different conceptual information, which could be adopted for genetic rule mining for lightweight ontologies. Currently, researchers from these two groups are collaborating to refine and visualise the generated lightweight ontologies. In this case study, we adopted these

two software tools to generate the lightweight ontology for European part of CIA FactBook.

CIA FactBook (http://www.odci.gov/cia/publications/factbook/index.html) is the input to the Corporum. Key concepts were extracted and the relations among them were identified by Corporum. The lightweight ontology about European countries was generated and represented in OIL (see Figure 3) and XML. The part of the manually-generated CIA FactBook by *Ontobroker* project was shown in Figure 4. Through the comparison, we can find some common concepts in both ontologies (see Figure 3 and 4). For instance, both of these two ontologies have identified that "waterway" has relation with "Transportation", one is defined as the property of "Transportation", while the other one specified the strength (as one of the slot-constraints) of this relation. This lightweight ontology can be further visualised by WebMaster.

```
Transportations[
   Railways =>>STRING;
   Highways =>>STRING;
   Waterways =>>STRING;
   Pipelines =>>STRING;
   Ports_and_harbors =>>STRING;
   Merchant_marine =>>STRING;
   Airports =>>STRING;
   Airports_with_paved_runways =>>STRING;
   Airports_with_unpaved_runways =>>STRING;
   Heliports =>>STRING].
```

Figure 4. Part of CIA FactBook ontology generated manually (Ontobroker, http://www.aifb.uni-karlsruhe.de/WBS/broker/wrapper/onto1.txt)

4. Related works

In IR, there are some researches having been done regarding to automatic thesauri ([6], [7]). Actually these IR researchers are already on their way to generate lightweight ontology (loosely speaking, they can be considered as lightweight ontologies, or linguistic ontologies (e.g. WordNet)).

Hwang [10] 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 seed-words, 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. The drawback is that it fully depends on the seedwords provided by the domain experts.

Maedche and Staab [11] proposed an approach to generate ontology semi-automatically based on the shallow text processing and learning algorithms. The dependency relations were extracted and treated as the input of the learning algorithms. Some of these relations didn't hold the meaningful relations of the two concepts linked together (co-occurrence) by some mediator (i.e., proposition). They also built up a system to facilitate the semi-automatic generation of the ontologies called Text-To-Onto. Kietz et al [12] adopted the above method to build an insurance ontology from a corporate Intranet.

Faure and Nedellec [13] presented an interactive machine learning system called ASIUM 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 [14] 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.

5. Discussion and future works

This paper illustrated the utilisation of co-occurrence theory to generate lightweight ontologies. The concepts (classes) of these lightweight ontologies were extracted from relevant domain documents. The relations of classes in these ontologies were embodied by either similarity in general or similarity with concrete values according to the co-occurrence theory. The subclass relations of the lightweight domain ontology were defined based on the broad-term or narrow-term relations in corresponding domain thesauri. These lightweight ontologies were represented in OIL.

In IR co-occurrence theory is a strategy for generation of thesauri semi-automatically or automatically (called automatic thesauri) which have the similar structure of these lightweight ontologies. While from AI point of view, these lightweight ontologies are just the start points of ontology learning. The explicit explanation power of various ontology representation languages, especially OIL, empowers these lightweight ontologies logical semantics and inference function. The contribution of these lightweight ontologies for ontology learning could be listed as follows:

- Ontology generating
  - The similarity relations in lightweight ontologies can be treated as the suggestion lists for some ontology editing tools (e.g. Protégé).
  - These lightweight ontologies could be treated as part of input for heavy-weight ontologies.
  - These lightweight ontologies could assist other efficient IR or IE expertises (e.g. hierarchy concept aligning) to generate fine-grained domain or task ontologies.
  - The similarity relations could become the starting point for relation-mining.
- Ontology mapping
  - The similarity measures of two classes (concepts) from different ontologies could provide valuable solution for the combination or integration of these ontologies.
  - The co-occurrence theory could cluster (group) similar classes (concepts) from different ontologies so as to generate the integrated ontologies (lightweight) (ontology clustering, see [15]).
- Ontology evolving
  - Dynamic changes of the similarity relations identified by co-occurrence theory provide good suggestions for evolving ontology (ontology maintenance) [16].
  - The co-occurrence theory could discover the associated relations of newly-merged concepts with other already-existing concepts [16]

For the future works, we will consider to use general linguistic ontologies (WordNet, Senses, Cyc) as the prototype to further refine the concepts and their relations identified by co-occurrence theory. We will also improve the method to visualise these lightweight ontologies. We will focus on the transformation or refinement of these lightweight ontologies to heavy-weight ontologies.

References

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

[2] Ingwersen, P. *Information retrieval interaction*. London: Taylor Graham, 1992.

[3] Stefik, M. *Knowledge Systems*. Morgan Kaufman Publishing, 1995.

[4] Berners-Lee, T. & Fischetti, M. *Weaving the Web*. Harper San Francisco, USA. 1999.

[5] Sheth, A. Semantic web and information brokering: Opportunities, early commercializations, and challenges. In *ECDL 2000 Workshop on the Semantic Web*, Sep 21, Lisbon Portugal, 2000.

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

[7] Chen, H., Martinez, J., Kirchhoff, A., Ng, T. D., & Schatz, B. R. Alleviating search uncertainty through concept associations. *Journal of the American Society for Information Science*, 49(3), 206-216, 1998.

[8] Noyons, E.C. M. & van Raan, A.F. J. Monitoring scientific developments from a dynamic perspective. *Journal of the American Society for Information Science*, 49(1), 68-81, 1998.

[9] Fensel, D. et al. OIL in a nutshell. In R. Dieng et al. (eds.), *Knowledge Acquisition, Modeling, and Lanagement, Proceedings of the European Knowledge Acquisition Conference (EKAW-2000)*, Lecture Notes in Artificial Intelligence, LNAI, Springer-Verlag, October 2000.

[10] 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.

[11] 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.

[12] 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, the International Conference in Grammere Inference (ICGI'2000) and Conference on Natural Language Learning (CoNLL'2000)*. Lisbon, Portugal, September 13-14, 2000.

[13] 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.

[14] 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.

[15] Visser, P.R.S. & Tamma, V.A.M. An experience with ontology clustering for information integration. In. *Proceedings of the IJCAI-99 Workshop on Intelligent Information*, Stockholm, Sweden, 1999.

[16] 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*, (2000), 26(6), 429-452.