Knowledge Acquisition for Semantic Search Systems

Wang Wei, Payam M. Barnaghi, Andrzej Bargiela
School of Computer Science
University of Nottingham (Malaysia Campus)
Jalan Broga, 43500 Semenyih, Selangor, Malaysia
{eyx6ww, payam.barnaghi, andrzej.bargiela}@nottingham.edu.my

Abstract

Semantic search extends the scope of conventional information search and retrieval paradigms from document-oriented and to entity and knowledge-centric search and retrieval. By attempting to provide direct and intuitive answers such systems alleviate information overload problem and reduce information seekers' cognitive overhead. Ontologies and knowledge bases are fundamental cornerstones in semantic search systems based on which sophisticated search mechanisms and efficient search services are designed. Nevertheless, acquisition of quality knowledge from heterogeneous sources on the Web is never a trivial task. Transformation of data in existing databases seems a promising bootstrapping approach, while information providers may refuse to do so because of intellectual property issues. In this article we discuss issues related to knowledge acquisition for semantic search systems. In particular, we discuss ontology learning from unstructured text corpus, which is an automatic knowledge acquisition process using different techniques.

1. Introduction

Variety of techniques have been developed by the information retrieval research community to facilitate retrieval of relevant documents in large text repositories [2]. Web search engines built on those techniques become indispensable tools for users to acquire useful information and knowledge in their day-to-day lives. To arrange more relevant documents on top of the result lists, most of contemporary Web search engines incorporate various ranking mechanisms (e.g., PageRank [5] and HITS [23]) that exploit link structures. Despite the substantial success, those search engines face perplexity in certain situations due to limitations such as superficial understanding of user queries and documents, and incapability of returning direct answers to the queries.

The semantic web [6] is based on the current Web in which resources are described using logic-based formal knowledge representation languages. The resource descriptions facilitate automated machine processing and agent cooperation across heterogeneous systems. In recent years, the semantic web technologies have been utilised to develop semantic search systems to supplement and improve conventional information retrieval systems. With traditional search systems, users have to inspect lists of URLs for documents and Web pages using their own knowledge in order to obtain satisfying answers, often by submitting multiple queries or combining information from different sources. On the contrary, semantic search systems extend scope of traditional information retrieval methods from document to entity and knowledge-centric methods. The latter supports providing direct and intuitive answers to alleviate the information overload problem and reduce users' cognitive overhead. Further, semantic search systems also enable design of search services to answer ad hoc queries, such as expert finding. Significance of the semantic search research has been demonstrated by number of existing semantic search systems. For example, entity and knowledge-centric search systems [16, 22, 18], semantic-enhanced question-answering systems [27], semantic association analysis [32, 1], mining-based search [34] (See [35] for a survey on semantic search systems).

Ontologies and knowledge bases serve as the fundamental building blocks based on which search, navigation and inference are performed, and to a great extent determine quality, coverage and usefulness of semantic search systems. In recent years, knowledge acquisition has attracted considerable amount of research attentions and various approaches and techniques have been proposed and utilised in this area. In this paper we cast the knowledge acquisition problem in the semantic Web as the task of ontology construction. Section 2 reviews types of ontologies in the semantic Web. Section 3 provides an overview of approaches and techniques solving the problem of knowledge acquisition. Section 4 is discusses the problem of ontology learning from unstructured text. Section 5 presents our ongoing research towards ontology learning using probabilistic topic models. Section 6 discusses the future work and provides a conclusion.

2. Types of Ontologies

In computer science, in particular in the context of semantic Web, ontologies provide formal conceptualisation of particular domains which are shared by a group of people. Sowa categorises ontologies into 3 types [1] :

- Formal ontology: a terminological ontology whose categories are distinguished by axioms and definitions stated in logic or in some computer-oriented language that could be automatically translated to logic.
- Prototype-based ontology: a terminological ontology whose categories are distinguished by typical instances or prototypes rather than by axioms and definitions in logic.
- Terminological ontology: an ontology whose categories need not be fully specified by axioms and definitions.

From the definitions one can see that a formal ontology directly defines mechanisms for logical inference to derive implicit knowledge; in a prototype-based ontology, categories are formed by collecting instances extensionally rather than describing the set of all possible instances in an intensional way [3]; in a terminological ontology, concepts are organised using subtype-supertype or part-whole relations, which corresponds to the "broader" and "narrower" relations in the SKOS vocabulary [2]. A well known example ontology of this kind is the ACM classification tree [3]. As mentioned earlier, we cast the knowledge acquisition as a task of ontology construction. As the unprecedented amount of information available on the Web highly limits the scale and viability of manual approaches to construct ontologies, we only focus on automated approaches towards ontology construction.

3. Knowledge Acquisition on the Semantic Web

Automated knowledge acquisition for the semantic Web, or ontology construction can be roughly classified into two classes: conversion-based approach and ontology learning from unstructured text. In most of the methods, a standard language processing procedure is performed, such as tokenisation, stop words elimination, part-of-speech tagging and stemming. We keep the main focus on ontology learning in the current paper and avoid further discussion on these procedures. For more information, reader can refer to [2].

Conversion-based approaches refer to those which transform structured (e.g., data in relational database) and semi-structured data (tables in Web pages) in machine processible format using formal knowledge representation languages, such as the RDF [4] and OWL [5]. TAP [16] is a semantic search system in which the knowledge base is populated by crawling and scraping Web pages; in RKB Explorer [15] and SWSE [18], the knowledge base is constructed by amalgamating information from a large number of different resources using a hybrid data-conversion solution, such as web page scraping, transformation of XML dump from DBLP into RDF; In ArnetMiner [34], a researcher profile knowledge base is populated against a research ontology using Conditional Random Fields [24], which is a probabilistic model for segmenting and labeling sequence data. It is worth mentioning that integration of data from different sources involves data consolidation and ontology mapping which are beyond scope of this paper. The advantages of using a conversion-based approach for ontology construction are large throughput and high accuracy. As such, the approach is suited for constructing formal ontologies against a pre-defined ontology schema. However, this approach also has some limitations: some domain specific knowledge might not be available in structured form; information providers might not choose to transform and expose the data due to proprietary concerns.

Ontologies can be learnt from various sources, be it databases, structured and unstructured documents or even existing preliminaries like dictionaries, taxonomies and directories [3]. Realising that wealth of the world knowledge is embedded in unstructured text on the Web, together with the assumption that given sufficient large amount of text in a domain, coverage of knowledge in that domain can be ensured, ontology learning has become a plausible solution for constructing ontologies out of unstructured text. Traditionally terminological and prototype-based ontologies are maintained manually, such as the ACM classification tree. The limitations can be identified: the manual approach is costly, time-consuming, and fast aging. On the contrary, ontology learning is a promising approach towards learning terminological ontologies (also referred to as concepts hierarchies [31, 29, 14, 37] by some researchers) as it is able to foster the conciseness of the model by determining meaningful and consistent generalizations [8]. The next section discusses different techniques and methods in ontol-

[1] http://www.jfsowa.com/ontology/gloss.htm

[2] http://www.w3.org/TR/swbp-skos-core-guide

[3] http://www.acm.org/class/1998/

[4] http://www.w3.org/TR/rdf-concepts/

[5] http://www.w3.org/TR/owl-guide/

ogy learning.

4. Ontology Learning from Unstructured Text

Cimiano [8] identifies six tasks for ontology learning from unstructured text organised as a "layered cake": terms, synonyms, concepts, concept hierarchies, relations, and rules. We classify the existing ontology learning methods in the literature into five categories based on the techniques deployed which are explained in the following.

4.1. Lexico-syntatic based Approach

Lexico-syntatic methods exploit regular expressions or repetitive patterns in natural languages. A well-known example is the Hearst-patterns [19] which is originally used to acquire hyponyms from large text corpora. For example, the following patterns help to identify that "France", "England", and "Spain" are hyponyms of "European Country":
NP {,} especially {NP,} * {or | and} NP (Original text: most European countries, especially France, England, and Spain.)
One of the weaknesses of this method is that Hearst patterns may not occur in the underlying texts. To address this issue, Cimiano et al [10, 9] utilise an method called "Learning by Googling" to exploit potential of large Web search engines to match such patterns for deriving super-sub concept relations and knowledge base population.

4.2. Information Extraction

Information Extraction (IE) [12] is a sub topic in natural language processing research. An IE application will take texts (structured or unstructured) as input and generate structured and unambiguous data. The simplest and most reliable IE technology is named entity recognition and has been deployed in some of the semantic search and ontology learning applications to automatically populate knowledge bases [22, 11]. The limitation is that the named entity recognition technology is domain-limited because it is only able to identify instances of general concepts such as "People", "Organisation", etc.

4.3. Clustering and Classification

Statistical machine learning techniques such as clustering and classification have also been adopted to learn ontologies out of unstructured text [28, 3]. An important underlying assumption for ontology learning from unstructured text is Harris' distributional hypothesis [17], which states that similar words tend to occur in similar contexts [3]. In clustering and classification-based ontology learning, terms are normally represented by those words in the vicinity of the represented terms.

Traditionally, clustering-based (e.g., hierarchical agglomerative/divisive, partitional) methods have been used to populate prototype-based ontology from scratch [7], while classification methods (e.g., K-nearest Neighbors [30]) have been used to augment a thesaurus with new lexical terms, that is, classifying new words to a large number of classes organised in a tree structure. In both of the approaches, similarity or divergence functions are important for the learning algorithms to determine to which clusters a new term or object is assigned. Various popular similarity and divergence functions have been employed and evaluated in existing research, for example, binary metric Jaccard's coefficient, geometric L1 norm, euclidean distance L2 norm, Cosine similarity measure and information-theoretic measures such as Kullback-Leibler divergence (KL) and Jensen-Shannon (JS) divergence (See formula 1 to 6) [26, 25]. The following gives the formulas for the above mentioned measures:

$$Jar(P,Q) = \frac{|\{P(i) > 0 \cap Q(i) > 0\}|}{|\{P(i) > 0 \cup Q(i) > 0\}|} \tag{1}$$

$$L1(P,Q) = \sum\nolimits_i |P(i) - Q(i)| \tag{2}$$

$$L2(P,Q) = \sqrt{\sum\nolimits_i (P(i) - Q(i))^2} \tag{3}$$

$$COS(P,Q) = \frac{\sum_i P(i) \bullet Q(i)}{\sqrt{\sum_i P(i)^2} \bullet \sqrt{\sum_i Q(i)^2}} \tag{4}$$

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{5}$$

$$D_{\mathrm{JS}}(P\|Q) = \frac{1}{2} D_{\mathrm{KL}}(P\|Q) + \frac{1}{2} D_{\mathrm{KL}}(Q\|P) \tag{6}$$

4.4. Probabilistic-based Approaches

Probabilistic Latent Semantic Indexing (PLSA) [21] and Latent Dirichlet Allocation (LDA) [4] are both probabilistic topic models which are originally developed for document classification and clustering in information retrieval community. The models are based upon the idea that instead of modeling relations between documents and words, documents are modeled as mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents which specifies a probabilistic procedure by which documents can be generated on the basis of latent variables. Its goal is to find the best set of latent variables that can explain the observed data [33].

In LDA, topics are extracted using Gibbs sampling algorithm which is a form of Markov Chain Monte Carlo [33].

In [37] the LDA learning processes are repeated with different number of topics. The learned topics are directly used as concepts contained in the underlying unstructured text corpora. Followed by a conditional independence test, a concept hierarchy with super-sub relations is derived in the domain of gene study. The author in [37] report higher precision, recall and F measure compared to other methods. The weakness of the method is the inability to infer subsumption relations in situations where a topic only subsumes one other topic, and identify some specific concepts. In an ongoing research, we use a dataset which contains around 5,000 abstracts of research articles in the semantic Web research domain. We apply the learned PLSA model (learning of the PLSA is through annealed Expectation-Maximisation algorithm [21]) and algorithms which iteratively compare similarity and divergence (e.g., Cosine similarity, KL, and JS divergence) between concepts to derive domain ontologies for the IRIS semantic search engine [36].

4.5. Data Co-occurrence Analysis

Another category of simple while effective methods for learning terminological ontologies is by processing the data co-occurrence. Sanderson et al [31] utilise a method based on an idea that a term A subsumes B if the documents in which B occurs are (or nearly) a subset of the documents in which A occurs. Despite simplicity of the idea, the experiment shows notable result compared to other methods based on Hearst patterns [19, 10]. In another work [29], the authors extend Sanderson et al's method [31] to represent each concept as a group of terms, and the subsumption relations between concepts are calculated using the subsumption relations of individual terms.

Another method which exploits co-occurrence of data is described in [14]. A variation of the PageRank [5] algorithm is utilised to exploit high-order data co-occurrence. The learned concept hierarchy is used in FacetedDBLP [6], which is a faceted browser that helps users to explore scientific publications.

5. Ontology Learning using PLSA

We are currently conducting a research on learning terminological ontologies using probabilistic topic models, in particular, probabilistic latent semantic analysis. The results of the ontology learning process will contribute to extend and enhance a semantic search system, IRIS [7]. PLSA is a statistic technique for the analysis of co-occurrence data [21]. It can be viewed as a probabilistic extension of the Latent Semantic Analysis (LSA) [13], which is a well-known technique for addressing the problems of "polysems" and "synonyms and semantically related words". In standard LSA, high dimensional word vector of documents are projected onto the lower dimension of latent semantic space using a technique called singular value decomposition of word-document tables [13]. The intuition is that by enfolding documents into lower dimensional space, semantically related documents are brought closer to each other. Compared to the LSA model, the PLSA has a sound statistical foundation and defines a proper generative model of data. In PLSA, unobserved latent variables, or topics, are introduced and associated to each observation. The underlying assumption of the model is that a document and a word are independent conditioned on the state of the associated latent variable [21]. Learning of parameters of the PLSA model is carried out using standard procedure for maximizing likelihood estimation using the Expectation-Maximisation (EM) algorithm (Hofmann introduced the annealed EM algorithm to avoid over-fitting [21]).

In most of the existing work, concepts are represented by their contextual information using simple words [19, 31], word phrases [14], or grammatical arguments [20, 7]. We use a piece of text, which is the centroid of documents describing the concepts calculated from our dataset. The idea is that a concept (e.g., a research topic in computer science) is a complex term and is better to be represented using a set of related contextual words. This representation of concepts fits well with the PLSA model:

- First, a PLSA model is learned using a dataset which consists of about 5,000 abstracts of scientific publication in the semantic Web research domain;

- Second, the concepts, which are in fact documents representing using vector of words with the "tf*idf" scheme [2], are projected onto the learned topic model (We select the concepts by counting the number of occurrence of the keywords annotating those documents). The resulting concepts are vector of hidden topics in the PLSA with lower dimension.

- Third, we use algorithms which iteratively calculate similarity and divergence values (e.g., Cosine similarity, KL, and JS divergence) between concepts to derive concept hierarchies automatically for the IRIS semantic search engine [36].

The left part of Figure 1 demonstrates the streamlined process of ontology learning process using PLSA and the right part shows part of the learned ontology.

The plausibility of the methods can be explained in two reasonable ways. From a traditional dimension reduction point of view, the learned PLSA model reduces the high dimension of word space to lower dimension of hidden topic space, thus semantically related concepts are brought

[6] http://dblp.l3s.de/

[7] http://rosie.nottingham.edu.my

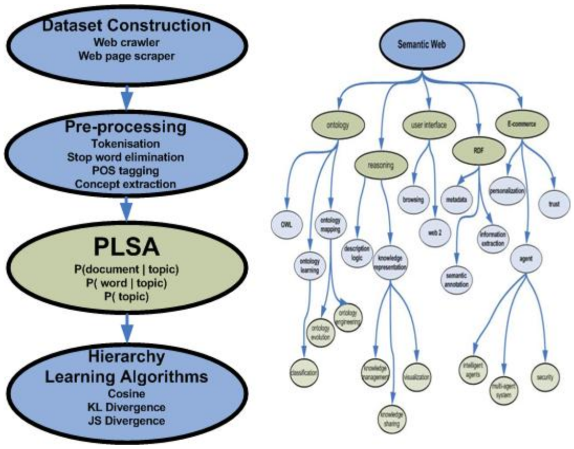

Figure 1. Components in ontology learning using PLSA and the learned ontology

closer. The second explanation is from a probabilistic point of view: after applying the PLSA model, the concepts are represented using a probability distribution of hidden topics which contains more "semantics" then raw words. The similarity or divergence between concepts is calculated based on probability distribution of topics. Concepts are similar or less divergent if they have a similar topic distribution.

6. Conclusion and Future Work

Automated knowledge acquisition for the semantic Web, or ontology learning has amalgamated considerable research efforts from communities such as natural language processing, machine learning, semantic Web, information retrieval, etc. Numerous methods have been proposed, implemented and evaluated in various systems. As reported in the literature, the average precision of the employed methods is not high enough to be used directly for formal knowledge representation and reasoning [8]. However, the automated ontology learning is useful for several purposes, for example, for applications where a certain error rate is tolerable, such as information retrieval, browsing and navigation [14]. Furthermore, in some of the existing ontology learning systems, the learned concepts are proposed to human ontology engineers for approval which significantly reduce the ontology engineering process.

The semantic Web community has produced a great number of ontology learning methods and techniques, nevertheless, these methods have not been implemented and deployed in large-scale knowledge-based systems except a few. The vision of the semantic Web is to enable people and computers to work in cooperation, which in a sense requires high-quality and reusable knowledge, i.e., ontologies. On one hand, the future work will involve improving the current existing methods and techniques to obtain better precision and high reusability, and applying novel techniques to enhance the existing approaches. On the other hand, the semantic Web also encourages design of novel applications which utilise those learned ontologies. Issues related to large-scale search applications, ranking and trust also need to be addressed. Moreover, due to the distributed and heterogeneous nature of the Web, efficient communications between software agents are difficult to establish. In such large environment where anyone is free to publish and consume information, reusing of knowledge and ontologies is proved to be extremely difficult. The linked data principle [8] provides reasonable guidelines for people to publish and link data on the Web. However, the adoption process is slow and not very effective due to absence of authorities. Research in ontology mapping partially alleviates the problem, while the community has not produced significant solutions and it remains as a future research question.

References

[1] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, L. Ding, P. Kolari, A. P. Sheth, I. B. Arpinar, A. Joshi, and T. Finin. Semantic analytics on social networks: experiences in addressing the problem of conflict of interest detection. In *WWW*, pages 407–416. ACM, 2006.

[2] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[3] C. Biemann. Ontology learning from text: A survey of methods. *LDV Forum*, 20(2):75–93, 2005.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[6] T. Burners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5), 2001.

[7] S. A. Caraballo. Automatic construction of a hypernym-labeled noun hierarchy from text. In *ACL*, 1999.

[8] P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[9] P. Cimiano, A. Pivk, L. Schmidt-Thieme, and S. Staab. Learning taxonomic relations from heterogeneous sources of evidence. In *Ontology Learning from Text: Methods, Evaluation and Applications*, Frontiers in Artificial Intelligence, pages 59–73. IOS Press, 2005.

[10] P. Cimiano and S. Staab. Learning by googling. *SIGKDD Explorations*, 6(2):24–33, 2004.

[11] P. Cimiano and J. Völker. Text2onto. In *NLDB*, pages 227–238, 2005.

[12] H. Cunningham. Information Extraction, Automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*, 2005.

[8] http://www.w3.org/DesignIssues/LinkedData.html

[13] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.

[14] J. Diederich and W.-T. Balke. The semantic growbag algorithm: Automatically deriving categorization systems. In *ECDL*, volume 4675, pages 1–13, 2007.

[15] H. Glaser and I. C. Millard. Rkb explorer: Application and infrastructure. In *Proceedings of Semantic Web Challenge*, 2007.

[16] R. V. Guha and R. McCool. Tap: A semantic web test-bed. *J. Web Sem.*, 1(1):81–87, 2003.

[17] Z. Harris. *Mathematical Structures of Language*. Wiley, 1968.

[18] A. Harth, A. Hogan, R. Delbru, J. Umbrich, S. ORiain, and S. Decker. Swse: Answers before links! In *Proceedings of Semantic Web Challenge*, 2007.

[19] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545, 1992.

[20] D. Hindle. Noun classification from predicate-argument structures. In *ACL*, pages 268–275, 1990.

[21] T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, pages 289–296, 1999.

[22] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *J. Web Semantics.*, 2(1):49–79, 2004.

[23] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA*, pages 668–677, 1998.

[24] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley and A. P. Danyluk, editors, *ICML*, pages 282–289. Morgan Kaufmann, 2001.

[25] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[26] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–, 1991.

[27] V. Lopez, M. Pasin, and E. Motta. Aqualog: An ontology-portable question answering system for the semantic web. In *ESWC*, pages 546–562, 2005.

[28] A. Maedche, V. Pekar, and S. Staab. Ontology learning part one - on discovering taxonomic relations from the web. *Web Intelligence*, pages 301–322, 2002.

[29] H. Njike-Fotzo and P. Gallinari. Learning generalization/specialization relations between concepts application for automatically building thematic document hierarchies. In *RIAO*, 2004.

[30] V. Pekar and S. Staab. Taxonomy learning - factoring the structure of a taxonomy into a semantic classification decision. In *COLING*, 2002.

[31] M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In *SIGIR*, pages 206–213, 1999.

[32] A. Sheth, I. B. Arpinar, and V. Kashyap. *Relationships at the Heart of Semantic Web: Modeling, Discovering, and Exploiting Complex Semantic Relationships*. Springer-Verlag, 2002.

[33] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2005.

[34] J. Tang, J. Zhang, D. Zhang, L. Yao, C. Zhu, and J. Li. Arnetminer: An expertise oriented search system for web community. In *Proceedings of ISWC2007*, 2007.

[35] W. Wang, P. M. Barnaghi, and A. Bargiela. A survey of semantic search systems. Technical report, School of Computer Science, University of Nottingham Malaysia Campus, 2008.

[36] W. Wei, P. M. Barnaghi, and A. Bargiela. The Anatomy and Design of A Semantic Search Engine. Technical report, School of Computer Science, University of Nottingham Malaysia Campus, 2007.

[37] E. Zavitsanos, G. Paliouras, G. A. Vouros, and S. Petridis. Discovering subsumption hierarchies of ontology concepts from text corpora. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 402–408, Washington, DC, USA, 2007. IEEE Computer Society.