

Search with Meanings: An Overview of 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
Email: {eyx6ww, payam.barnaghi, andrzej.bargiela}@nottingham.edu.my

Abstract: Research on semantic search aims to improve conventional information search and retrieval methods, and facilitate information acquisition, processing, storage and retrieval on the semantic web. The past ten years have seen a number of implemented semantic search systems and various proposed frameworks. A comprehensive survey is needed to gain an overall view of current research trends in this field. We have investigated a number of pilot projects and corresponding practical systems focusing on their objectives, methodologies and most distinctive characteristics. In this paper, we report our study and findings based on which a generalised semantic search framework is formalised. Further, we describe issues with regards to future research in this area.

Keywords: Semantic Search, Knowledge Acquisition, Semantic Web, Information Retrieval.

1. Introduction

Research in information retrieval (IR) community has developed variety of techniques to help people locate relevant information in large document repositories. Besides classical IR models (i.e., Vector Space and Probabilistic Models) [6], extended models such as Latent Semantic Indexing [17], Machine Learning based models (i.e., Neural Network, Symbolic Learning, and Genetic Algorithm based models) [14] and Probabilistic Latent Semantic Analysis (PLSA) [29] have been devised with hope to improve information retrieval process. However, rapid expansion of the Web and growing wealth of information pose increasing difficulties to retrieve information efficiently on the Web. To arrange more relevant results on top of the retrieved sets, most of contemporary Web search engines utilise various ranking algorithms such as PageRank [11], HITS [32], and Citation Indexing [33] that exploit link structures to rank the search results. Despite the substantial success, those search engines face perplexity in certain situations due to the information overload problem on one hand, and superficial understanding of user queries and documents on the other.

The semantic web [12] is an extension of the current Web in which resources are described using logic-based knowledge representation languages for automated machine processing across heterogeneous systems. In recent years, its related technologies have been adopted to develop semantic-enhanced search systems. Significance of the research in this area is clear for two reasons: it supplements conventional information retrieval by providing search services centered on entities, relations, and knowledge; and development of the semantic web also demands enhanced search paradigms in order to facilitate acquisition, processing, storage, and retrieval of the semantic

information. The paper provides a survey to gain an overall view of the current research status. We classify our studied systems into several categories according to their most distinctive features, as discussed in the next section. The categorisation by no means prevents a system from being classified into other categories. Further, we limit the scope of the survey to Web and Intranet searching and browsing systems (also including some question answering and multimedia presentation generation systems). Hildebrand *et al* discuss the related research from three perspectives: query construction, search algorithm and presentation of results [28]. We provide a review focusing on objectives, methodologies, and most distinctive features of individual systems; and discuss issues related to knowledge acquisition and search methodologies. The rest of the paper is organised as follows. Section 2 discusses the studied work. Section 3 proposes and formalises a semantic search framework. Section 4 describes the future work in this research area and concludes the paper.

2. Semantic Search Systems

Conventional search techniques are developed on the basis of words computation model and enhanced by the link analysis. On one hand, semantic search extends the scope of traditional information retrieval paradigms from mere document retrieval to entity and knowledge retrieval; on the other hand, it improves the conventional IR methods by looking at a different perspective: the meaning of words, which can be formalised and represented in machine processible format using ontology languages such as RDF¹ and OWL². For example, an arbitrary resource or entity can be described as an instance of a class in an ontology; having attribute values and relations with other entities. With the logical representation of resources, a semantic search system is able to retrieve meaningful results by drawing inference on the query and knowledge base. As a simple example, meaning of the query for “people in School of Computer Science” will be interpreted by a semantic search system as individuals (e.g., academic stuff) who have relations (e.g., work for or affiliated) with the school. On the contrary, conventional IR systems interpret the query based on its lexical form. Web pages in which the worlds “people” and “computer science” co-occur, are probably retrieved. The cost is that users have to extract useful information from a number of pages, possibly query the search engine several times. As explained in the following, other inference mechanisms based on logical rules and inductive approaches have also been evaluated to enable a system to interpret and understand ad-hoc queries.

¹ <http://www.w3.org/TR/rdf-schema/>

² <http://www.w3.org/TR/owl-guide/>

We classify the research of semantic search into six categories in accordance with their objectives, methodologies, and functionalities. Categorisation and discussion of semantic search systems based on alternative criteria such as user-system interaction can be found in [34].

2.1 Document-oriented Search

Document-oriented search can be thought of as an extension of the conventional IR approaches where the primary goal is to retrieve documents, such as web pages and text documents [27, 38, 31], scientific publications [48, 19] and ontologies [20]. In search systems of this category, documents are annotated using logic-based knowledge representation languages with respect to annotation and domain ontologies to provide approximate representations for the topics or contents of original documents. The retrieval process is carried out by matching user queries with the resulting semantic annotations.

The early work in SHOE search system facilitates users constructing constrained logical queries by specifying attribute values of ontology classes to retrieve precise results [27]. The main limitations are the manual annotation process and lack of inference support. There has been considerable work on speculating the significance of logical reasoning for semantic search systems. OWLIR [38] adopts an integrated approach that combines logical inference and traditional information retrieval techniques. A document is represented using original text and semantic markup. The semantic markup and the domain ontology are exploited by logical reasoners to provide enhanced search results. Conventional IR system based on the original text is integrated into the semantic search system to provide complementary results in case no result is returned by the semantic search. In a similar work, a semantic search framework for annotation, indexing and retrieval of documents called KIM is introduced by Kiryakov *et al* [31]. The automated annotation framework is based on information extraction technologies concerning named entities. With the semantic annotation of the content, constrained queries with regard to entity type, name, attribute and relation are formulated to obtain precise results. In addition, a query pre-processing step with logical inference enables the system to interpret user queries in meaningful ways [31]. One of the common limitations of the above discussed work is that the context of development is based on a “closed world” assumption that does not take heterogeneous nature of the semantic web into consideration (e.g., sources may publish different ontologies in similar domains).

Several semantic search systems have also been developed to provide alternative ways for researchers to explore and browse large repositories of scholarly articles [48, 18]. The IRIS [48] provides an inference engine to perform reasoning with rules over a computer science literature domain ontology to elicit implicit knowledge. During user interaction with the system, semantically related, broader and narrower concepts (the IRIS domain ontology is built on SKOS³) are recommended to facilitate user browsing and refining queries. FacetedDBLP [19] is a faceted browser that helps users to browse scientific publications based a number of facets. The concept facet is built using the GrowBag algorithm which exploits keywords co-occurrences to construct concept hierarchies [18].

A semantic web document is different from a standard document (e.g., web page, email and scholarly article) because it is authored primarily for automatic machine processing, though it is human readable. Further, unlike hyperlinks between Web documents, links between semantic web documents are labeled with meanings. To cope with the problems, an ontology ranking algorithm which is based on a “rational surfer model” rather than “random surfer model”, OntoRank, is introduced in [21]. An search engine called Swoogle [20] is implemented to search ontology documents on the Web.

2.2 Entity and Knowledge-oriented Search

Entity and knowledge-oriented search methods expand the scope of conventional IR systems which solely retrieve documents. Systems based on this method often model ontologies as directed graphs and exploit links between entities to provide exploration and navigation. Attribute values and relations of the retrieved entities are also shown to provide additional knowledge and rich user experiences [23, 25, 42, 22].

TAP is one of the first empirical studies of large scale semantic search on the Web. It improves the traditional text search by understanding the denotation of the query terms, and augmenting search results using a broad-coverage knowledge base with an inference mechanism based on graph traversal [23]. Recently a number of systems such as CS-Aktive [42], RKB Explorer [22], and SWSE [25] have been implemented based on the principles of “open world” and “linked data⁴”, which coincide with spirit of the semantic web of being “distributed”. CS-Aktive and RKB Explorer provide unified views of information (using graphical interfaces) collected from a significant number of heterogeneous data sources. To resolve the problem that heterogeneous sources may publish different information about same set of entities, CS-Aktive uses its reference ontology as the mediator [42], while RKB Explorer implements set of consistent reference services, which are essentially knowledge bases of URI equivalence generated using heuristics [22]. A notable feature of the SWSE system is a hybrid data integration solution for the collected data, such as existing RDF datasets, XML database dumps, various static and live data sources. Data consolidation is realised using a simple mechanism through analysis of inverse functional properties of entities [25]. Retrieved results are ranked by using a modified PageRank algorithm using context information (i.e., provenance of data) [30].

2.3 Multimedia Information Search

In ontology-based image annotation and search systems, resources are annotated by domain experts or using text around images in documents. Early works adopt manual approach which suppresses scalability [41]. The work in [47] performs query expansion to retrieve semantically related images by drawing logical inference on its domain ontology (e.g., using subsumption relation between concepts). Falcon-S [49] annotates images using metadata by crawling and parsing pages on the Web. Disambiguation of resources with same labels is resolved using context information derived from user queries [49]. Squiggle [13] is a semantic framework to help building domain-specific semantic search applications for indexing

³ <http://www.w3.org/TR/swbp-skos-core-guide>

⁴ <http://www.w3.org/DesignIssues/LinkedData.html>

and retrieving multimedia items. Its knowledge representation model is based on the SKOS vocabulary which enables the system to suggest meanings of queries by a simple inference process, e.g., suggest alternative labels or synonyms for an image. As a result, images annotated with one label can be retrieved using the image's alternative labels.

2.4 Relation-centered Search

Relation-centered semantic search approach pays special attention to relations between query terms implicitly expressed by users. It usually performs an additional query pre-processing step through the use of external language lexicons, or an inference process using a knowledge base [37, 34, 31, 35].

AquaLog is a question-answering system. It implements similarity services for relations and classes to identify synonyms of verbs and nouns appearing in a query using WordNet⁵. The obtained synonyms are used to match property and entity names in the knowledge base for answering queries [37]. SemSearch supports natural language queries and translates them into formal queries for reasoning [34]. An entity referred by a keyword is matched against a subject, predicate or object in the knowledge base using combinations.

The above two systems process entity relations using language lexicon or word combinations. While in KIM [31] and OntoLook [35], relations between query terms are inferred from knowledge base to aid the retrieval process, for example, in KIM a query “company, Redwood Shores” could be understood by the system that the intention of the query is to retrieve documents mentioning the town and companies with geographical constraints (i.e., companies located in the town), but not the word “company” [31]. OntoLook assembles query terms to concept pairs and send these pairs to the knowledge base to retrieve all relations which have been asserted in the knowledge base.

2.5 Semantic Analytics

Semantic analytics, also known as semantic association analysis, is introduced by Sheth *et al.* to discover new insights and actionable knowledge from large amounts of heterogeneous content [43]. It is essentially a graph-theoretic based approach that represents, discovers and interprets complex relationships between resources. By modeling RDF database as directed graph, relationships between entities in knowledge base are represented as graph paths consisting of a sequence of links.

Search algorithms for semantic associations such as breadth first and heuristics-based search are discussed in [44]. Effective ranking algorithms are indispensable because the number of relationships between entities in a knowledge base might be much larger than the number of entities. In [2] the ranking evaluates a number of parameters: context, subsumption, trust, rarity, popularity, and association length. In another work called SemRank, a blend of semantic and information-theoretic techniques are used to analyse and rank the semantic associations [4]. Semantic association analysis has been deployed in several real world applications, such as national security application [44] and conflict of interest detection [3]. A relation robustness evaluation algorithm for semantic associations has been implemented in the MANA multimedia presentation generation sys-

tem to automatically select multimedia objects for presentation generation [7].

2.6 Mining-based Search

Effectiveness of the semantic search depends largely on the quality and coverage of the underlying knowledge base. The search methodologies discussed so far either utilise explicit knowledge, which is asserted in the knowledge base, or implicit knowledge, which is derived using logical inference with rules. Another kind of knowledge, which we refer to as “hidden knowledge”, can not be easily observed using techniques such as information extraction, natural language processing, logical inference, and semantic analytics. For example, “who are the experts in semantic web research community?”, “Which institutions ranks highly in the machine learning research area?”. Such knowledge can only be derived from large amount of data by using some sort of sophisticated data analysis techniques. We refer to approaches that utilise techniques to infer hidden knowledge as mining-based semantic search.

Flink is a semantic web application for extraction, aggregation and visualization of an online social network that consists of professional work and social connectivity of semantic web researchers [39]. It allows users to identify prominent researchers in the semantic web community based on popular measures in social network analysis. Another system called Ontocopi uses a method that combines breath first and spreading activation search to identify communities of practice from its knowledge base [1]. Arnetminer [46] is another semantic-enhanced application which is featured by a number of mining services, e.g., expert finding. Using researchers' publication, the expert mining service utilises the PLSA model to extract and rank experts with respect to user queries.

3. Formalisation of A Semantic Search Framework

The objective to optimise search results has motivated research in the semantic search area by incorporating techniques from variety of other research fields and implementation of a number of practical systems. However, our investigation of the existing work reveals that a formalised framework is not introduced by any of the existing works which covers most aspects of a semantic-enhanced information search and retrieval process.

The work in [27, 38, 34, 37] only describes different components of their respective systems, [31] introduce a “closed world” semantic search architecture in which distributed and heterogeneous nature of the information sources is not considered. The recent work in [42, 25, 22, 8] is developed based on the open world assumption and linked data principle. The knowledge acquisition problem, in particular, information integration and consolidation from different sources has been emphasised while search mechanisms have been paid less attention. Moreover, in the architecture of the systems, demarcation between different functionality components is not clearly defined; modules providing similar functionalities across systems are positioned in different components with different names. The framework presented in [46] is limited in its specific application settings and needs to be generalised to accommodate other related research methodologies. Therefore, there is a need for a formalised framework to accommodate and consoli-

⁵ <http://wordnet.princeton.edu/>

date existing frameworks while easily extensible and adaptable to new application requirements.

3.1 Overview of the Framework

We propose such a framework as shown in Figure 1. The framework is logically abstracted to six distinctive components responsible for semantic data acquisition, knowledge acquisition, data integration and consolidation, semantic search mechanisms, semantic search services, and result presentation. Formalisation of such a framework addresses some of the problems presented in the existing work: the framework identifies the common tasks that a semantic search system should implement according to the six components; each component unambiguously defines functionalities should be offered by that component and provides requisite input to the one on top, compared to the existing semantic search systems where division between functionalities are not clearly defined; each processing component can be added or removed in accordance with requirement of the specific application and availability of data. In Figure 1 we also summarise prevalent methodologies and techniques for each component which have been proposed and adopted in the existing work.

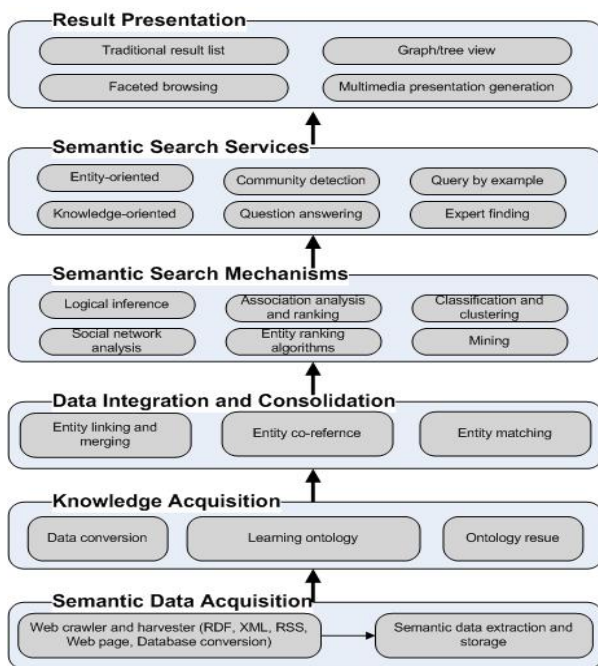


Fig. 1. A semantic search framework

The data acquisition component provides different solutions for collecting unstructured (e.g., web pages), semi-structured (e.g., data in XML and database), and structured semantic data (e.g., existing RDF datasets and RSS feeds). The collected unstructured and semi-structured data needs to be transformed to structured data using certain processing techniques defined in the knowledge acquisition component.

Methods used for acquiring knowledge in semantic Web applications fall into mainly two categories: conversion-based methods and ontology learning (we skip the discussion of ontology reuse issue here). Conversion-based approaches refer to those transform semi-unstructured data into structured data according to pre-defined schemas. It is effective when large amount of semi-structured data and well defined schemas are

available. However, there are some problems associated with approaches of this kind which prevent them being efficiently deployed. For example, conversion-based approaches do not provide broad coverage of some domains; in some application domains semi-structured data may not be available or data publishers are not willing to publish such data due to proprietary concerns. Moreover, approaches of this class are not able to generate structured data from unstructured data. Research in ontology learning [31, 26, 9, 40, 18, 15] provides promising solutions for generate structured data, or ontologies, from unstructured text with tolerable error rates. The assumption is that given sufficient large amount of text in a domain, coverage of knowledge in that domain can be ensured. As a consequence, the problem of “knowledge acquisition bottleneck” [15] can be alleviated to a great extent. For example, hidden knowledge, which is essential for applications such as expert finding and concept hierarchy construction while is difficult to observe, could be obtained using some sophisticated inference techniques in an automated fashion. In section 3.2 we discuss various techniques for learning ontologies from unstructured text and outline our current research for learning ontologies using probabilistic topic models [29, 10, 45].

The data integration and consolidation component summarises solutions for a problem arisen from the knowledge acquisition process, which is the fact that different sources may publish different and often complementary data on same entities. The “linked data” principle provides a number of guidelines (e.g., entity linking and co-referencing) for web sites to avoid publishing same data repeatedly. Further, reconciliation of heterogeneous semantic entities generated based on different ontologies needs ontology and entity matching. In section 3.3 we discuss some of the major difficulties of the current methods and proposals towards the entity consolidation problem and propose a method for that problem based on the idea of entity grouping using clustering techniques.

We distinguish between search mechanisms and services: by search mechanisms we mean various techniques based on which semantic search services are implemented, such as social network analysis [39], semantic association analysis [43, 44, 3], and PLSA [29, 50]; while the search services provide an abstract model of the functionalities a semantic search engine offers. Semantic search services have extended the scope of services that conventional IR techniques could offer, such as entity and knowledge search. Further, existing applications such as question answering and expert finding, which have been studied using conventional IR methods, could be also improved by semantic technologies [37, 50]. The framework streamlines a semantic search system to a set of autonomous while related processes whose functionalities are clearly defined. Design of one process is independent of another, for example, different knowledge base, different search mechanisms and ranking algorithms can be defined.

In [28] the authors have provided a detailed discussion regarding result presentation and we will not repeat the subject due to the space limitation. In what follows we focus on discussion of ontology learning for knowledge acquisition and entity consolidation.

3.2 Ontology Learning from Unstructured Text for Knowledge Acquisition

Ontology learning from text has attracted much attention in the semantic Web research community in recent years due to the “knowledge acquisition bottleneck” problem [15]. Existing approaches can be classified into four categories: Lexico-syntactic based approach [26, 15], Information Extraction [16, 31], Clustering and Classification [9], and Data Co-occurrence Analysis [40, 18].

Lexico-syntactic methods exploit regular expressions and repetitive patterns in natural languages. A well-known example is the Hearst-patterns which can be used to acquire hyponyms from large text corpora [26]. The weakness is that desired patterns may appear infrequently or not appear in the text corpora, resulting in a low recall rate. Information Extraction techniques, in particular named entity recognition, have also been used to automatically populate knowledge bases [31]. The limitation is that the named entity recognition is domain-limited because it is only able to identify instances of general concepts such as “People”, “Organisation”, etc. Traditionally, ontology learning methods based on clustering techniques have been used to populate prototype-based ontology from scratch. However, the limitation is that automatic labeling of the learned clusters remains a problem. Methods based on classification have been used to augment a thesaurus with new lexical terms while the main problem is that human labeling of training data is required [9]. Data Co-occurrence Analysis is a simple while effective method for learning terminological ontologies is by exploiting data co-occurrence in text collection. Sanderson *et al* [40] introduce a method based on the assumption 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. An experiment shows notable result compared to other methods, such as Lexico-syntactic methods. Another method utilising co-occurrence of data is described in [18] which uses an algorithm called the semantic GrowBag. The learned ontology (i.e., concept hierarchy) is used in FacetedDBLP⁶ browser to help users exploring scientific publications.

In an ongoing research we employ probabilistic topic models, namely, Probabilistic Latent Semantic Analysis (PLSA) [29] and Latent Dirichlet Allocation (LDA) [10] to learn terminological ontologies for the IRIS semantic search engine⁷. PLSA and its enhanced descent, LDA are probabilistic topic models [45] for the analysis of co-occurrence data. Both bring in the models latent variables called topics and define probabilistic generative processes for words and documents in the text corpora. A document and a word are independent conditioned on the state of the associated latent variable. A document is generated by first choosing a distribution over topics; then for each word in that document one “chooses a topic at random according to that distribution, and draws a word from that topic” [45] (LDA is an improved model of PLSA in which conjugate Dirichlet priors are added for document-topic and topic-word multinomial distributions). Learning of parameters is carried out using inference based on maximizing likelihood principle estimation using the Expectation-Maximisation (EM) algorithm [29] in PLSA. The inference technique used in LDA is Gibbs Sampling, one of the Markov Chain Monte Carlo al-

gorithms [45] (This approach is different from the algorithm of learning parameters described in the original LDA paper [10]).

We represent concepts (i.e., research topics in Computer Science) using centroid of documents which describe those concepts from our dataset. The assumption is that a concept is a complex term and is better to be represented using a set of related contextual words which is in line with Harris’ distributional hypothesis [24]. The procedure of applying the PLSA model for learning concept hierarchy is summarised as follows. (also see Figure 2. Figure 3 shows parts of the learned concept hierarchy):

- A PLSA model is learned using a dataset which consists of about 5,000 abstracts of scientific publication in the semantic Web research domain;
- The concepts, which are in fact documents representing using vector of words with the “tf*idf” scheme [6], are projected onto the learned topic model. The resulting concepts are vectors of hidden topics with lower dimension.
- The algorithms which iteratively calculate similarity and divergence measures (e.g., Cosine similarity, KL, and JS divergence [36]) between concepts are used to derive concept hierarchies automatically.

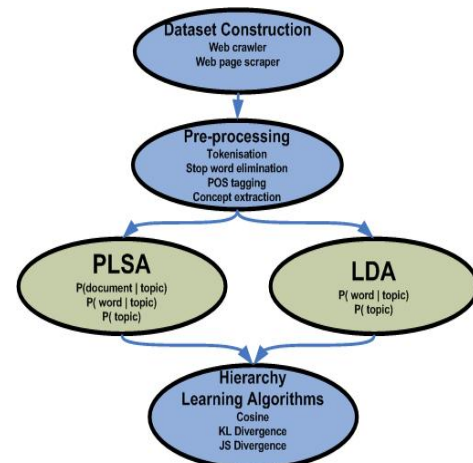


Fig. 2. The concept hierarchy learning procedure

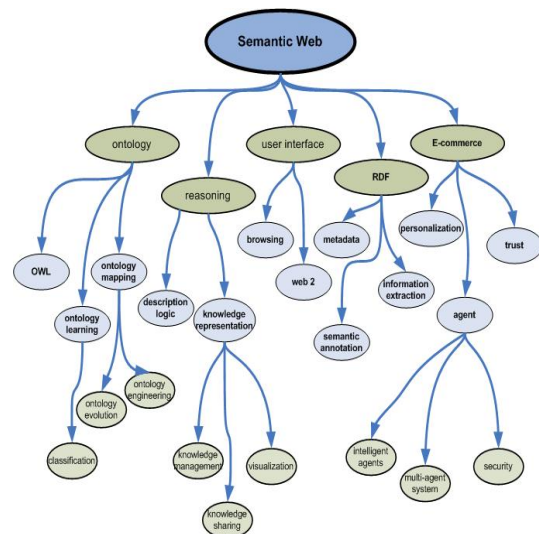


Fig. 3. A fragment of concept hierarchy in semantic Web research constructed using learning algorithms

⁶ <http://dblp.l3s.de/>

⁷ <http://rosie.nottingham.edu.my>



Fig. 4. Result returned by SWSE search engine

3.3 Entity Consolidation

In this sub-section we discuss one of the specific issues which is related to entity consolidation in an open world semantic search system.

The distributed nature of the Web allows different sources to publish different and often complementary information on same entities. The problem lies in that these sources often duplicate “multiple copies” of an entity. As a consequence, semantic search engines are likely to return a list of URLs of the entities which in fact are the same (e.g., repetitive information) and normally complete information about an entity is scattered by a list of URLs (e.g., complementary information). Users still need to investigate the URL list one by one and use their prior knowledge to determine what information to trust even in the presence of ranking algorithms. Further, since one source is unlikely to publish complete information on an entity of interest, users often have to combine information gathered from multiple sources. In this scenario, compared to conventional search engines, the user cognitive overhead has not been reduced as much as expected. A snapshot of the result from SWSE search engine using query “Tim Berners-Lee” is shown in Figure 4. The left part is a list of all classes of the instance denoted by the query. The right part is a list of URLs about the query.

The above described situation contradicts with the original intention of semantic search systems, which aims to find answers for users by directly returning information or knowledge on entities in a efficient manner. The current major solution proposed for this problem is the “linked data principle”. However it is difficult to be implemented due to the unprecedented scale of the Web and lack of coordinating authorities. Ontology mapping is another solution, however, the research mainly concentrates on schema level and does not provide sufficient support at instance level. We propose the use of partitional clustering techniques to group those entities which in fact refer to the same object. It is essentially a pre-processing step for semantic search engines before answering user queries. Given data collected from the Web (e.g., using semantic crawlers) and the distributional hypothesis, a contextual vector is constructed as representation of an entity (e.g., using words describe the entities or words describe other entities which directly linked to the entity). Alternatively, the vector of an entity is represented using its attributes. Partitional clustering algorithms are then used to group duplicated copies of entities (they may carry comple-

ment information) based on standard similarity measures in the underlying knowledge base before being deployed in the semantic search engine. Despite the simplicity of the idea, it is able to alleviate the problems outlined earlier.

4. Conclusion and Future Work

Semantic search is a joint discipline that brings together research from communities of IR, semantic web, machine learning, natural language processing, information extraction, and so on. it aims to expand the scope and improve retrieval quality of conventional IR techniques. We have investigated a number of existing systems, and classified them into several categories in accordance with their methodologies, scope, and functionalities. The main finding is: though varieties of systems have been developed, a logical semantic search framework is not formalised. We have proposed an extensible and adaptable framework addressing common tasks and issues in the related research, and provided a detailed discussion on two of the components in the framework, i.e., ontology learning and entity consolidation.

Ontology and knowledge base are fundamental cornerstones for designing useful semantic search services. However, knowledge acquisition is a bottleneck for semantic-enhanced applications. Because conversion of tremendous amount of unstructured data into structured data is not a trivial task, research on automated knowledge acquisition, in particular, ontology learning [15], have attracted much attention in recent years. Cimiano *et al.* categorise ontology learning into several subtasks including concept hierarchy induction, learning attributes and relations, and ontology population [15]. An array of automated techniques to learning ontologies out of large text corpus have also been proposed. Although these methods have achieved success to a certain extent, there is much space for the learning accuracy to be improved.

The future work is also concerned with the trust and quality of knowledge. In the Web where anyone is free to publishing data, the quality of the knowledge varies largely from source to source. Effective ranking algorithms are needed to distill most trustworthy and quality information. Because of the lack of a universal representation for entities (e.g., in conventional IR systems, documents are represented by vector of term weights) on the semantic web, trust is one of the most substantial parameters in semantic search systems. A survey of the current trust models for the semantic web is provided in [5]. Further, new evaluation metrics need to be devised for semantic search systems. Large scale experiments and user studies are also required to be carried out to assess effectiveness of implemented systems.

References

- [1] H Alani, K O’Hara and N Shadbolt, Ontocopi: Methods and tools for identifying communities of practice. Intelligent Information Processing 2002, Vol. 221, pp. 225–236.
- [2] B Aleman-Meza, C Halaschek-Wiener, I B Arpinar, C Ramakrishnan and A P Sheth, Ranking complex relationships on the semantic web. IEEE Internet Computing, Vol. 9, No. 3, 2005, pp. 37–44.
- [3] 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.. WWW 2006, ACM, pp. 407–416.

- [4] K Anyanwu, A Maduko and A P Sheth, Semrank: ranking complex relationship search results on the semantic web. WWW 2005, ACM, pp. 117–127.
- [5] D Artz and Y Gil, A survey of trust in computer science and the semantic web. J. Web Sem., Vol. 5, No. 2, 2007, pp. 58–71.
- [6] R A Baeza-Yates and B A Ribeiro-Neto, Modern Information Retrieval, ACM Press / Addison-Wesley, 1999.
- [7] P M Barnaghi and S A Kareem, Relation robustness evaluation for the semantic associations. EJKM, Vol. 5, 2007, pp. 265–272.
- [8] A D L Battista, N Villanueva-Rosales, M Palenychka and M Dumontier, Smart: A web-based, ontology-driven, semantic web query answering application. Proceedings of ISWC2007 2007.
- [9] C Biemann, Ontology learning from text: A survey of methods. LDV Forum, Vol. 20, No. 2, 2005, pp. 75–93.
- [10] D M Blei, A Y Ng and M I Jordan, Latent dirichlet allocation. Journal of Machine Learning Research, Vol. 3, 2003, pp. 993–1022.
- [11] S Brin and L Page, The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, Vol. 30, No. 1–7, 1998, pp. 107–117.
- [12] T Burners-Lee, J Hendler and O Lassila, The semantic web. Scientific American, Vol. 284, No. 5, 2001.
- [13] I Celino, E D Valle, D Cerzza and A Turati, Squiggle: a semantic search engine for indexing and retrieval of multimedia content. Proceedings of SAMT 2006 2006, pp. 20–34.
- [14] H Chen, Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms.. JASIS, Vol. 46, No. 3, 1995, pp. 194–216.
- [15] P Cimiano, Ontology Learning and Population from Text: Algorithms, Evaluation and Applications, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [16] H Cunningham, Information Extraction, Automatic. Encyclopedia of Language and Linguistics, 2nd Edition, 2005.
- [17] S C Deerwester, S T Dumais, T K Landauer, G W Furnas and R A Harshman, Indexing by latent semantic analysis. JASIS, Vol. 41, No. 6, 1990, pp. 391–407.
- [18] J Diederich and W T Balke, The semantic growbag algorithm: Automatically deriving categorization systems.. ECDL 2007, Vol. 4675, pp. 1–13.
- [19] J Diederich, W T Balke and U Thaden, Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp. JCDL 2007, ACM, p. 505.
- [20] L Ding, T Finin, A Joshi, R Pan, R S Cost, Y Peng, P Reddivari, V Doshi and J Sachs, Swoogle: a search and metadata engine for the semantic web. CIKM '04 New York, NY, USA, 2004, pp. 652–659.
- [21] L Ding, T W Finin, A Joshi, Y Peng, R Pan and P Reddivari, Search on the semantic web. IEEE Computer, Vol. 38, No. 10, 2005, pp. 62–69.
- [22] H Glaser and I C Millard, Rkb explorer: Application and infrastructure. Proceedings of Semantic Web Challenge 2007.
- [23] R V Guha, R McCool and E Miller, Semantic search. WWW 2003, pp. 700–709.
- [24] Z Harris, Mathematical Structures of Language, Wiley, 1968.
- [25] A Harth, A Hogan, R Delbru, J Umbrich, S Oriain and S Decker, Swse: Answers before links!. Proceedings of Semantic Web Challenge 2007.
- [26] M A Hearst, Automatic acquisition of hyponyms from large text corpora. COLING 1992, pp. 539–545.
- [27] J Heflin and J Hendler, Searching the web with SHOE. Artificial Intelligence for Web Search Menlo Park, CA, 2000, pp. 35–40.
- [28] M Hildebrand, J van Ossenbruggen and L Hardman, An Analysis of Search-based User Interaction on the Semantic Web, Tech. rep., Centrum voor Wiskunde en Informatica (NL), 2007.
- [29] T Hofmann, Probabilistic latent semantic analysis. UAI 1999, pp. 289–296.
- [30] A Hogan, A Harth and S Decker, Reconrank: A scalable ranking method for semantic web with context. Proceedings of SSWS 2006.
- [31] A Kiryakov, B Popov, I Terziev, D Manov and D Ognyanoff, Semantic annotation, indexing, and retrieval. J. Web Semantics., Vol. 2, No. 1, 2004, pp. 49–79.
- [32] J M Kleinberg, Authoritative sources in a hyperlinked environment. SODA 1998, pp. 668–677.
- [33] S Lawrence, C L Giles and K Bollacker, Digital libraries and Autonomous Citation Indexing. IEEE Computer, Vol. 32, No. 6, 1999, pp. 67–71.
- [34] Y Lei, V S Uren and E Motta, Semsearch: A search engine for the semantic web. EKAW 2006, pp. 238–245.
- [35] Y Li, Y Wang and X Huang, A relation-based search engine in semantic web.. IEEE Trans. Knowl. Data Eng., Vol. 19, No. 2, 2007, pp. 273–282.
- [36] J Lin, Divergence measures based on the shannon entropy.. IEEE Transactions on Information Theory, Vol. 37, No. 1, 1991, pp. 145–.
- [37] V Lopez, M Pasin and E Motta, Aqualog: An ontology-portable question answering system for the semantic web. ESWC 2005, pp. 546–562.
- [38] J Mayfield and T Finin, Information retrieval on the semantic web: Integrating inference and retrieval. Proceedings of Workshop on Semantic Web at SIGIR 2003.
- [39] P Mika, Flink: Semantic web technology for the extraction and analysis of social networks. J. Web Semantics, Vol. 3, No. 2-3, 2005, pp. 211–223.
- [40] M Sanderson and W B Croft, Deriving concept hierarchies from text. SIGIR 1999, pp. 206–213.
- [41] A T Schreiber, B Dubbeldam, J Wielemaker and B J Wielinga, Ontology-based photo annotation. IEEE Intelligent Systems, Vol. 16, No. 3, 2001, pp. 66–74.
- [42] N Shadbolt, N Gibbins, H Glaser, S Harris and M M C Schraefel, Cs active space, or how we learned to stop worrying and love the semantic web. IEEE Intelligent Systems, Vol. 19, No. 3, 2004, pp. 41–47.
- [43] 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.
- [44] A P Sheth, B Aleman-Meza, I B Arpinar, C Bertram, Y Warke, C Ramakrishanan, C Halaschek, K Anyanwu, D Avant, F S Arpinar and K Kochut, Semantic association identification and knowledge discovery for national security applications. J. Database Manag., Vol. 16, No. 1, 2005, pp. 33–53.
- [45] M Steyvers and T Griffiths, Probabilistic topic models, In Latent Semantic Analysis: A Road to Meaning, T Landauer, D Mcnamara, S Dennis, and W Kintsch, Eds. Laurence Erlbaum, 2005.
- [46] J Tang, J Zhang, D Zhang, L Yao, C Zhu and J Li, Arnetminer: An expertise oriented search system for web community. Proceedings of ISWC 2007.
- [47] W Wei and P M Barnaghi, Semantic support for medical image search and retrieval. BIEN '07 Anaheim, CA, USA, 2007, pp. 315–319.
- [48] W Wei, P M Barnaghi and A Bargiela, The Anatomy and Design of A Semantic Search Engine, Tech. rep., School of Computer Science, University of Nottingham Malaysia Campus, 2007.
- [49] H H Wu, G Cheng and Y Z Qu, Falcon-s: A ontology-based approach to searching objects and images in the soccer domain. ISWC 2006.
- [50] J Zhang, J Tang, L Liu and J Li, A mixture model for expert finding. To Appear in Proceedings of 2008 Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2008) 2008.