Chapter 1

State of the Art

1.1 Introduction

Social media has become an important source of knowledge, user-generated content has a great potential useful data in terms of business opportunities and research data source. In this dissertation, the author performs a case study on Linked.com, a leading websites in Social Media, and build a knowledge model for the company and professional public profiles. The potential use of the dataset could be similar to [1], where applications will be built on top on the dataset to provide the user with customised data aggregation.

This project focuses on developing the knowledge model representations of user generated content in the context of Semantic Web. Semantic Web can be regarded as an revolution from Web of documents to Web of data and knowledge.[2] The key factor that differentiate it from traditional web is that, it guarantees machine-readable data that supports automatic reasoning. It increase the interoperability of the data by defining the semantic meanings. Normally, The Resource Description Framework(RDF) is used to describe the resources.

In the context of the Semantic Web, there's a movement that tries to encourage information holders to publish and link their data together; it's called Linked Data. More and more people contribute to the Linked Data Cloud[3], for example, government Linked

Data has already been maintained by W3C.org, Ontologies and RDF are heavily used in Biomedical domain, the FOAF project has already attracted Social networks to use it to model the users, and the DBpedia, the Semantic version of the Wikipedia, has become the centre of the Web Ontoloies[4]. So we decide to build our knowledge model using RDF, because it can take the advantages of Semantic Web, to support reasoning and machine auto-processing. Apart from that, SPARQL Protocol and RDF Query Language(SPARQL), can be used to infer the facts from RDF triples.

This project focuses on developing the knowledge model representations of user generated content in the context of Semantic Web. Ideally, the data model should be general enough so that new knowledge can be inferred from the extracted data. Because we are using RDF triples to represent data, SPARQL will be used as the query tool to answer questions.

In order to generate knowledge models from raw HyperText Markup Language (html) files of LinkedIn public profiles, a number of challenges are required to be addressed, such as Data Extraction, Knowledge Modelling, Content Integration and Evaluation of Extracted Data. In the next section, we discuss each challenge in detail.

1.2 Data Extraction

1.2.1 Data extraction in general

[5] provides an up-to-date survey on web data extraction. In this paper, three common techniques for web data extraction is listed: 1. Tree-based approach: analysis on Document Object Model (DOM) trees. 2. Web wrapper: use procedures to seeks and finds data required. 3. Machine learning approach: using reasoning or other Artificial Intelligence (AI) techniques to find the data of interest. In addition, the paper provides a full list of famous applications that are being used in the real world. In our approach, as we can only access to the HTML files of LinkedIn public profiles, Web wrapper method

will be used to extract data. Although the pages do not contain structure knowledge, the format are consistent and barely change. Even some profiles are incomplete, we can handle this in our Wrapper program.

[6] discusses four challenges or concerns that every research will encounter in the field of Semantic Web and Big Data.

1. Michael L. Brodie mainly focuses on data integration. He also provides a general form for it: 1. Define the concern. 2. Search for candidate data elements. 3. Extract, transform and load (ETL) the candidate data into appropriate formats. 4. Entity resolution to get unique, comprehensive data. 5. Answer the query/solve the problem.

2. Christian Bizer tries to motivate people to take the Billion Triple Challenge (`http://challenge.semanticweb.org/`). The challenge is about using pre-crawled data set to translate different vocabularies into uniform one, discover resources and fuse descriptions into an integrated representation. So the main challenges here are: 1. Large-scale RDF processing. 2. Data quality. 3. Data Integration.

3. Peter Boncz proposes the Linked Open Data Ripper, a web portal to combine open government data. The main challenges are the accessibility and the usability of the public government data. He is looking for robust, reliable user interfaces that integrate Linked Data from multiple sources and allow users to query the data more easily.

4. Orri Erling believes systematic adoption of Database Management System (DBMS) technology into Semantic Web could be a potential opportunity, since efficient storage and query of DBMS has been researching for decades. A lot of optimisation mechanisms, performance tools have been developed to support the system. The challenges exist are: 1. we need to demonstrate the benefit of semantics. 2. smarter database is required for reasoning, but Web Ontology Language (OWL) is not

enough. 3. we need to bring Linked data and RDF into the regular data-engineering stack.

These challenges are interesting topics that waiting to be addressed. Nevertheless, it provides a brief overview of the current status of Big Data stack.

[7] gives a relative short introduction of several ways to mine data from LinkedIn.com, typically, LinkedIn Search, raw data processing, and third-party tools. Among these approaches and tools, the Python Natural Language Toolkit (NTLK) and [8] are two resources that worth to study.

1.2.2 Data Extraction approaches

[9] approaches the problem of web table data extraction by using two-dimensional visual box model. This paper introduces extracting information from a high level of visual features. It uses the representation of web browser rendering, and save the practitioner from parsing low level CSS, JavaScript, HTML tags. The key difference is that, the traditional approach uses tree-based representation of web pages (HTML/XML), so the whole information extraction is processed in low level, using HTML/XML parses. As far as the author can tell, this approach only works for tables and lists, so it cannot be applied to arbitrary elements on web pages.

[10] discuss about automatically extracting concepts from semi-structured data, specifically, they use PowerPoint slides as the knowledge source. They combine ontology learning and natural language processing techniques to produce the knowledge representation. The process as follows: 1. normalising the text contents by splitting statements, replacing non-alphanumeric symbols, expanding abbreviations, etc. 2. creating parse tree for sentences. 3. defining a set of weighting models. 4. Extracting text features (e.g. topic, title, bullet, sub bullet) for each term and applying "link-distance algorithm" to determine to correct concepts. What can be learned from the paper is that they effectively use Natural Language Processing to tag each term and then define weighting models to hierarchically

extract concepts using text features. But the problem still exists, that is, the 42% of overall performance (F-measure) is not enough to apply this techniques into real world E-learning application. Apart from that, in their future work, they plan to introduce multi-media feature extraction into the their paper. The author believes the high values of F-measure is very important for real use of this technique, which is the thing that this paper cannot handle.

[11] presents a framework that exploits the Web documents using a "Tree Alignment Algorithm", in which they build trees iteratively and try to find record boundary and repeating patterns. Then they build "conceptual graphs" to represent domain knowledge. Finally they map the conceptual structure to the extracted data items. Because the conceptual graph is directly mapped to a database schema, this approach can reduce the time of converting the extracted content to database records. The approach proposed here could be very useful in this project, which also trying to extract data of interest from semi-structure LinkedIn profile files. However, as far as the author can tell, the approach might be not scalable, as manually creating a "conceptual graph" is required, which makes the approach no better than using pure "Regular Expression" approach. Nevertheless, we can learn from the "mapping" process and adopt it. In our case, Levenshtein distance (Edit distance) or Cosine similarity (Vector space model) could be used to classify vocabularies and correct typos.

[12] describes a method to populate Wikipedia info-boxes from Wikipedia article. It trains "value extractors" from training data using structural analysis. Structure discovery algorithm is used to overcome the shortcomings of regular expression, in which it tries to merge important patterns from a frequent pattern list. One thing is not clear in this paper is how to choose correct attribute value among a list of potential attribute values. It does mention using "Conditional Random Fields" (CRF) to learn label tokens based on features. "Combining regular expressions" provides better results, it worths further investigation.

[13] talks about metadata extraction from enterprise content. It performs a case study

on documents that described by Docbook DTD, which is used widely by many organisations. The motivation of the paper is to provide a novel framework for personalised information retrieval system. It also generate an Ontology for user modelling. This approach is deeply couple with the Docbook content, similar approach might be used in this project as our data are deeply couple with LinkedIn html structure.

1.3 Knowledge Modelling

We are living in the era of Web 2.0, which means that large scale of the user-generated content are available on the Internet in a loose or semi-structure format. Traditionally, Data Mining is performed on relational databases or data warehouse, in a way that practitioners look for unaware patterns internally. But gathering data from blooming Social Media websites cannot be fully addressed in the traditional approach, as most of the websites are producing HTML or XML files. A mapping between raw data format and relational database table is required but hard to generalise to other data consumers.

That's why we need Semantic Web. Semantic Web aims to replace the Web of documents to the Web of machine processable, automatic reasoning web services or web databases. It provides interoperability to data by strictly narrow down the data into triples and allow each piece reference others using unique resource identifier. The potential of Semantic Web is difficult to estimate, it might totally change the development paradigm[14]: data drive possible applications instead of what we do today, applications determine data format.

1.3.1 Semantic Web Technologies

Semantic Web technology stack can be thought as a layered graph as shown in Figure 1.1. We are going to introduce some of the technologies in this section.

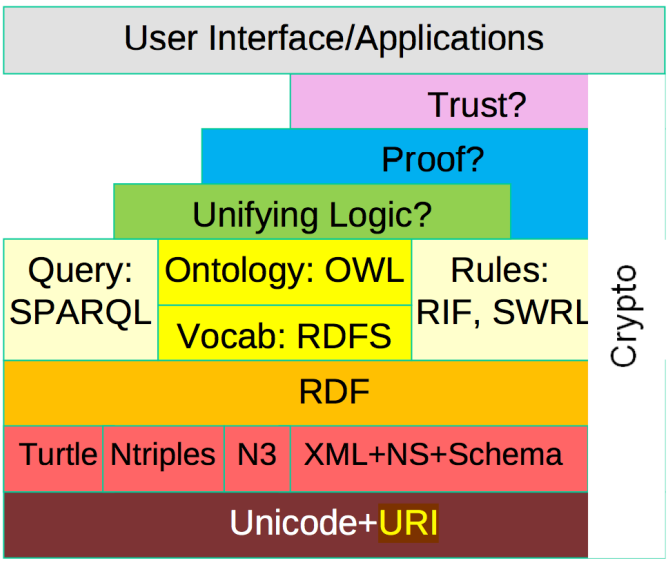

Figure 1.1: Semantic Web Stack[1]

Uniform resource identifier (URI) : A string that can be used to uniquely identify a web resource. According to [15], URI is considered as the standard resource identifier to represent any HTML or RDF object or concept. The reason behind that is others can easily access the resources using Hypertext Transfer Protocol (HTTP) requests.

Resource Description Framework (RDF) : RDF is a graph-based data model that used in Semantic Web. It represents knowledge using a triple structure. An expression in RDF is a "subject-predicate-object" triple. It can be represented by a graph where the subject and the object is the start node and end node and the predicate is the link. Nodes can be a URI or a literal. It has a variety of notations such as N3, Turtle, and XML, but they are all interchangeable. Notice that it just a data model that allows us to describe things that in a specific syntax, but has no assumption.

RDF Schema (RDFS) : It intends to provide vocabularies to standardise the structure of RDF resources. It is a set of classes and properties that use the RDF language to provide basic ontologies. The reason we need these vocabularies is that RDF Triple itself is not informative enough. Different datasets need a standard (just like

[1]Copy from Dr. Rob Brennan's lecture notes.

protocol) to communicate, otherwise, no one will understand the "semantic" of other datasets. So RDFS specifies a basic vocabulary such as: subClassOf, DataType, domain, range, etc. to structure the RDF resources.

Web Ontology Language (OWL) : OWL aims to add more constraints on RDFS to describe resources in details. For example, owl defines disjointWith, complementOf, equivalentClass, and cardinality on top of RDFS. It makes the triple expression become more expressive and specific.

Link Open Data Movement : It's a community effort starting in 2006 to unlock hidden semantics in a way that making RDF publicly available using open standards and protocols. Many open datasets were published by these efforts from many domains such as geographic locations (e.g. Geonames), general knowledge (e.g. DBpedia), broadcasting data (e.g. BBC), bioscience, etc. The best way to feel the impact of the Cloud is to visualise the graph (Figure 1.2):

[16] demonstrates how to collect, analyse FOAF documents. According to the paper, FOAF is one of the most popular ontology that being used at the moment. One of the main produces of FOAF documents is blog website. It's easy to use FOAF specific tags to identify the documents, and looking for patterns. Apart from above, the reader known from it that LinkedIn.com also use the FOAF ontology, but they protect the FOAF documents from public access. This paper implies that we can use FOAF Ontology to describe LinkedIn public profiles and extend it if necessary.

[3] provides a comprehensive state of the art on the Linked Open Data (LOD). It introduces "Linked data principles", how to publishing Linked Data, publishing tools, existing applications. Also, related developments and research challenges are given to guide the later researchers.

[2]Attribution: "Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/"

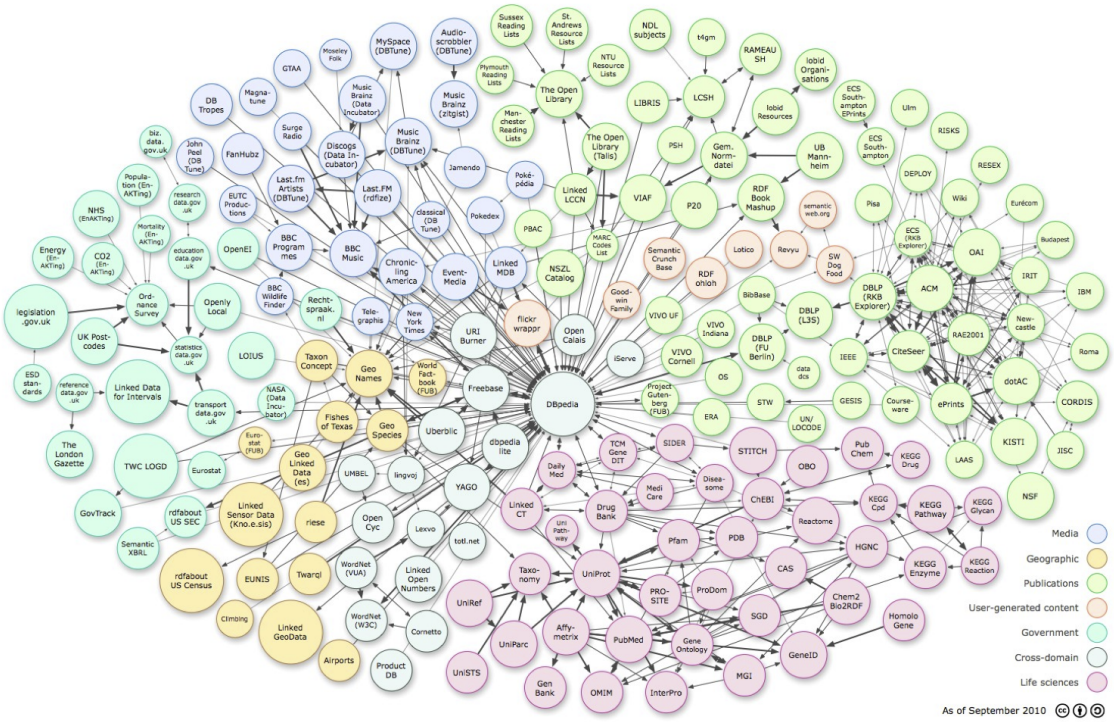

Figure 1.2: Link Data Cloud[2]

1.3.2 Linked Open Data

As mention in the previous section, Linked Open Data is a movement that data providers start to publish and link their data to each other in RDF format. It enables[17]: sophisticated data processing, connecting distributed data and change the world from Data Islands to a Global Data Space. The bar for publishing the Linked data is not restrict, as it only needs to conform following basic principles[15]:

1. Use URIs to name the things and use HTTP URIs to guarantee accessibility.
2. Provide standard information when users access it (RDF*, SPARQL).
3. Include links to other URIs.

As more and more important websites join in this movement, LOD is becoming a huge knowledge graph as shown in Figure 1.2. DBpedia, a RDF version of Wikipedia,

now become the centre of Open Data Cloud. A lot of tools have been built for LOD[3], for instance, Linked Data search engines, allow you quickly look for the documents you are interested in; Publishing tools, allow you quickly publish the data in RDF even though the origin files in display in other format. Our project aims to contribute to the LOD in a way that we provide queryable web service that allow people to discover important facts and statistics.

[4] provides a brief introduction about the DBpedia. It firsts talk about the extraction of structure information from Wikipedia, which is followed by a list of datasets. Finally, it talks about how to access, query the dataset online (using HTTP, SPARQL endpoint and RDF dump), how DBpedia interlink with other open datasets, and how to search DBpedia.org using built-in user interface. Through the paper, the authors try to convey a fact that DBpedia is the nucleus of the Web of Data, which is a reasonable claim.

[18] talks about a live extraction framework that can consume Wikipedia updates and reflect on the DBpedia triple store. The key process is as follows: 1. Use different extractors to deal with different types of content. 2. Assign states to extractors, namely, an extractor could be in either "updated", "not modified", or "remove" state. 3. Apply heuristic method (by comparing current Axiom to previous one) to minimise the number of triples that need to be updated. To increase the effectiveness of "mapping" between Wikipedia and DBpedia, templates are introduced to infer the correct attribute names and correct values. Keeping DBpedia content up-to-date has several benefits, such as enhancing the integration with Wikipedia, increase the use of DBpedia in live scenarios. So later if the project want to keep the Ontology and triples up-to-date and reflect the instant change in LinkedIn.com, using the approach mention in this paper could be a potential solution.

[19] gives a detailed introduction of DBpedia Spotlight – a Web Service to detect DBpedia resources in text. The key improvement of the disambiguation process is: instead of using traditional "TF-IDF" to weight the words, it uses "TF-ICF" (term frequency-Inverse Candidate Frequency). Moreover, to maximise the annotation result, the authors

suggest use customised configuration when annotating. This web service could be very useful when later the reader tries to annotation the data fields in LinkedIn public profiles.

1.3.3 Building ontologies

We can think of it as a collection of terms that defines the concepts and relationships of an area[3]. It is the cornerstone of the Semantic Web; by publishing ontologies and combining them together, the web of knowledge will finally be constructed.

[20] mainly focus on the strategy of building simple Ontologies for social networks. A tripartite model is suggested in this paper, specifically, an Actor-Concept-Instance model. The paper demonstrates the applicability of the model using two examples. The paper also shows how the ontology is emerged based on the model and how it is extended to support Ontology Extraction from Web Pages. However, this approach mainly about Community Ontology Construction, as LinkedIn public profiles has no or very limit connection information. In our approach, we will try to enhance linkage/mapping to other datasets, like DBpedia (for general information), Academic Institution Internal Structure Ontology (AIISO) (for academic skills, courses), Freebase (for general subjects), etc.

[21] focuses on extracting information from Artificial Intelligence related conference and workshop and building an Ontology for AI. Again, it constructs domain concept knowledge from nested tags. for example, in HTML, <h1>means a more general term than <h2>, so an instance of <h2>is a subclass of an instance of <h1>. Then in the optimisation process, it performs "ontology pruning and union" to handle concept duplication. However, this strategy might result in wrong classification. To summarise, this approach is very useful provided the user knows the contents in the web pages is valid for hierarchical classification. It could not be generalised for other loose structured websites.

In this project, our goal is to build an Ontology for LinkedIn public profiles using automated process. The reasons for doing that are, firstly, Linkedin.com is one of the

[3]http://www.w3.org/standards/semanticweb/ontology

main knowledge sources for professional information. People publish their education, skills, experiences on the site and we expect these kinds of information can answer a lot questions. For example, decision maker may want to track the trending of an industry by looking at the number of employees and the number of new startups in the specific area. Secondly, we choose Semantic Web because we want to link the knowledge into the Web of knowledge (LOD) to maximise the usability of our data. The interoperability feature provided by RDF can lead to flexible use of triples (Again, in this case, data can drive the application developments)

1.4 Content Integration and Classification

One of the major problems in information extraction (IE), especially in social media information extraction is the variety of the similar words. For example, in LinkedIn.com, a user can claim himself as "Graduated from Trinity College Dublin", meanwhile, another user will say she is studding at "TCD". When we build an ontology and try to link our data to LOD, we really have to be very careful about declaiming a term more than once. A false positive result is also not acceptable, in a way that we might misclassify address "Dublin" in "Dublin Core" as the capital of Ireland. So finding ways to clean up the data and classify them correctly are considered two complex tasks in IE.

[22] gives a very comprehensive introduction about machine learning in text categorisation. document indexing and dimensionality reduction are common techniques to increase the effectiveness of accessing data. Probability classifiers, decision tree classifiers, on-line methods, neural networks, etc. At last, measures of effectiveness was discussed. At the moment, we will not try to parse the "Summary" section in public profiles (In LinkedIn profiles, the summary section is where users write "abstract" about themselves). But if we need more detail knowledge for the Ontology, we might use the approach listed in this paper.

[23] proposes an approach to build re-usable dictionary repositories for text mining.

The key idea is to build a new dictionary by using synonyms from existing dictionary. They only use synonym relations, which cannot be enough to represent more complex semantic relations. And if the practitioner choose an inappropriate dictionary to start with, he will end up getting nothing back since the similarity value is too low. Apart from that, according to the authors, the idea of generating text corpus for the existing dictionaries can save about 50%-60% of time.

[24] talks about how to integrate government data from different data sources. The integration flow is as follows: 1. Mapping and Scrubbing. They maps attributes to a simple global schema, and cleansing on data value level. 2. Data Transformation, in which they transforms the source data structure to the global schema and separates data of different types. 3. Deduplication. A tool called Duplicate Detection Toolkit was used to match across data sources. 4. Entity Fusion. They fused the matched entities to obtain a single representation. "Dempster-Shafer-Theory" is used to induce weights for attributes. We can investigate the mapping process since we will require map person to other linked dataset instances.

1.5 Evaluation of the Extracted Data

As everyone can publish their data on the Internet, the evaluation of the data quality becomes a very important aspect in Ontology building. People cannot or hard to reuse the data with bad quality, so publishing the data without quality assurance will significantly reduce the value and the reusability of the data. Therefore, we evaluated some metrics and a data assessment framework:

[25] lists quality metrics for metadata. This project can use some of these metrics directly to evaluate the quality of the result of the data mining. Data Completeness is achieved by comparing extracted data with the "ideal" representation. Data Accuracy can be achieved by the degree of correctness. In this project, it's possible to compare manually collected data with the automatically extracted data. We can use user studies,

by introduce volunteers to extract the data. Then by investigating the manually collected results to machine auto-generated ones, we will have some confidence about our data correctness. Conformance to expectations is a way to test whether the schema meets the requirement of use cases, and supports arbitrary complex queries. Because our dataset will be used by another project: "Leveraging Power of Social Media and Data Visualisation", we can evaluate the dataset by looking at whether the data is complied with the user and visualisation requirements. So, the metrics list in this paper can evaluate the quality of the data.

[26] presents a Linked data quality assessment and fusion framework that can be used to measure, express the quality of data. It's a part of the Linked Data Integration Framework (LDIF). The integration process works as follows: 1. access web data, 2, map the vocabulary from different schema using R2R framework. 3. LDIF also resolutes multiple identifiers for the same entity by using "Silk-Link Specification Language". 4. the data quality assessment module contains a set of scoring functions, and it also support user-extend scoring function and customisation. 5. finally, the data fusion module includes conflict ignoring, avoiding and resolution strategies to "sieve" the data and generate a cleaner representation. Since this paper focus on both quality measurement and data fusion, what we can use from this paper is the Data Quality Assessment module. It's possible to use the built-in scoring functions directly or implement new methods.

Bibliography

1. Li, Y., Shi, Y., Fan, X. & Bhavsar, M. *CareerGalaxy A planner for future* 2012.

2. Shadbolt, N., Hall, W. & Berners-Lee, T. The Semantic Web Revisited. *Intelligent Systems, IEEE* **21,** 96–101. ISSN: 1541-1672 (Jan.-Feb.).

3. Bizer, C., Heath, T. & Berners-Lee, T. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)* **5,** 1–22 (2009).

4. Auer, S. *et al.* Dbpedia: A nucleus for a web of open data. *The Semantic Web,* 722–735 (2007).

5. Ferrara, E., De Meo, P., Fiumara, G. & Baumgartner, R. Web Data Extraction, Applications and Techniques: A Survey. *arXiv preprint arXiv:1207.0246* (2012).

6. Bizer, C., Boncz, P., Brodie, M. & Erling, O. The meaningful use of big data: four perspectives–four challenges. *ACM SIGMOD Record* **40,** 56–60 (2012).

7. Bradbury, D. Data mining with LinkedIn. *Computer Fraud & Security* **2011,** 5 –8. ISSN: 1361-3723 (2011).

8. Russell, M. *Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites* (O'Reilly Media, 2011).

9. Gatterbauer, W., Bohunsky, P., Herzog, M., Krüpl, B. & Pollak, B. in *Proceedings of the 16th international conference on World Wide Web* (2007), 71–80.

10. Atapattu, T., Falkner, K. & Falkner, N. in *Database and Expert Systems Applications* (2012), 161–175.

11. Hemnani, A. & Bressan, S. Extracting information from semi-structured Web documents. *Advances in Object-Oriented Information Systems,* 389–396 (2002).

12. Lange, D., Böhm, C. & Naumann, F. in *Proceedings of the 19th ACM international conference on Information and knowledge management* (2010), 1661–1664.

13. Sah, M. & Wade, V. in *Proceedings of the 19th ACM international conference on Information and knowledge management* (ACM, Toronto, ON, Canada, 2010), 1665–1668. ISBN: 978-1-4503-0099-5. doi:10.1145/1871437.1871699. ¡http://doi.acm.org/10.1145/1871437.1871699¿.

14. Bergman, M. Advantages and Myths of RDF. *AI3, April* (2009).

15. Berners-Lee, T. *Design issues: Linked data* 2006.

16. Ding, L., Zhou, L., Finin, T. & Joshi, A. in *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04* (IEEE Computer Society, 2005), 113.3–. ISBN: 0-7695-2268-8-4. doi:10.1109/HICSS.2005.299. ¡http://dx.doi.org/10.1109/HICSS.2005.299¿.

17. Heath, T. & Bizer, C. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology* **1,** 1–136 (2011).

18. Hellmann, S., Stadler, C., Lehmann, J. & Auer, S. in *Proceedings of the Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009 on On the Move to Meaningful Internet Systems: Part II* (Springer-Verlag, Vilamoura, Portugal, 2009), 1209–1223. ISBN: 978-3-642-05150-0. doi:10.1007/978-3-642-05151-7_33. ¡http://dx.doi.org/10.1007/978-3-642-05151-7_33¿.

19. Mendes, P. N., Jakob, M., García-Silva, A. & Bizer, C. in *Proceedings of the 7th International Conference on Semantic Systems* (ACM, Graz, Austria, 2011), 1–8. ISBN: 978-1-4503-0621-8. doi:10.1145/2063518.2063519. ¡http://doi.acm.org/10.1145/2063518.2063519¿.

20. Mika, P. Ontologies are us: A unified model of social networks and semantics. *Web Semant.* **5,** 5–15. ISSN: 1570-8268 (Mar. 2007).

21. Wang, S., Zeng, Y. & Zhong, N. Ontology extraction and integration from semi-structured data. *Active Media Technology,* 39–48 (2011).

22. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv.* **34,** 1–47. ISSN: 0360-0300 (Mar. 2002).

23. godbole, S., Bhattacharya, I., Gupta, A. & Verma, A. in *Proceedings of the 19th ACM international conference on Information and knowledge management* (ACM, Toronto, ON, Canada, 2010), 1189–1198. ISBN: 978-1-4503-0099-5. doi:10.1145/1871437.1871588. ¡http://doi.acm.org/10.1145/1871437.1871588¿.

24. Bohm, C. *et al.* in *Proceedings of the 6th International Conference on Semantic Systems* (2010), 34.

25. Ochoa, X. & Duval, E. in *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2006* (eds Pearson, E. & Bohman, P.) (AACE, Chesapeake, VA, 2006), 1004–1011. ¡http://www.editlib.org/p/23127¿.

26. Mendes, P., Mühleisen, H. & Bizer, C. in *Proceedings of the 2012 Joint EDBT/ICDT Workshops* (2012), 116–123.