Combining Linked Data and Statistical Information Retrieval

Next Generation Information Systems

Ricardo Usbeck

University of Leipzig, Germany
{usbeck|ngonga}@informatik.uni-leipzig.de
R & D, Unister GmbH, Leipzig, Germany

Advisors: Axel-Cyrille Ngonga Ngomo, Andreas Both and Sören Auer

Abstract. Being a part of the *Information Age*, users are challenged with a tremendously growing amount of Web data which generates a need for more sophisticated information retrieval systems. The *Semantic Web* provides necessary procedures to augment the highly unstructured Web with suitable metadata in order to leverage search quality and user experience. In this article, we will outline an approach for creating a web-scale, precise and efficient information system capable of understanding keyword, entity and natural language queries. By using Semantic Web methods and *Linked Data* the doctoral work will present how the underlying knowledge is created and elaborated searches can be performed on top.

Keywords: Search, NLP, Question Answering, Ranking

1 Introduction

In the last couple of years, the way search is perceived by end users as well as industrial agents changed dramatically. Recently, new semantic search algorithms[1] spread which account not only for keywords but for semantic entities, relations, personalized information and many more. In analogy, future developments in everyday and business search engines need to unlock the power of semantic technologies.

Linked Data is the Semantic Web methodology for publishing data based on W3C standards such as RDF [22], URI and HTTP in order to provide linkable, valuable content. Whether provided by a SPARQL [2] endpoint or embedded in a Web page via RDFa [3], Linked Data is a key technology to master the upcoming information flood. Since 2007, the Linked Open Data (LOD) Cloud gathered more than 300 datasets also known as *knowledge bases* comprising over

[1] `http://searchengineland.com/google-hummingbird-172816`

31 billion triples[2]. Amongst others it consists of agricultural, musical, medical and geographical facts, the LOD Cloud is the largest linked encyclopaedic knowledge base known to mankind.

Using the Semantic Web is expected to drive innovation in data integration and analysis software within companies. Moreover, end users anticipate more sophisticated search engines that truly understand the underlying information need. Therefore, combining scientifically sound information retrieval methods with static and dynamic Web data as well as Linked Data will leverage information insight already in the short term. For example, fundamental scientific work has been done in the Linked Open Data [4] project. However, there is no information retrieval framework which is able to convert the scientific knowledge into a holistic Semantic Web-based search engine.

In Section 2, the state of the art in the areas of information retrieval and Linked Data-based search and ranking algorithms is presented. The problems tackled in this thesis and its contributions are described in Section 3. Section 4 presents the already available approaches *AGDISTIS* [34], which is a named entity extraction framework for unstructured Web pages, and *REX* [6], a relation extraction approach for templated websites. Furthermore, first steps towards an auto-completion functionality are pointed out and plans on further research regarding search and ranking algorithms are presented. Section 5 concludes with an outlook on the future research agenda.

2 State of the Art

(1) *Information Extraction.* This field can be considered as comprising three main sub-fields: named entity recognition (NER), named entity disambiguation (NED) and relation extraction (RE). NER is the task of identifying entities in an input text while NED is focused on pre-identified named entities and their disambiguation towards a certain knowledge base using various methods. RE is the task of finding connections between entities based on a given context. In this thesis, we restrict the identifiable entity classes to 'persons', 'locations' and 'organizations' using FOX [26] as well-known NER framework.
In the following, several NED approaches for *unstructured texts* are introduced. A framework for annotating and disambiguating Semantic Web resources in unstructured texts is DBpedia Spotlight [24]. Contrary to other tools, Spotlight is able to disambiguate against all classes of the DBpedia ontology. Another algorithm is AIDA which uses the YAGO2[3] Linked Data knowledge base using sophisticated sub-graph matching algorithms. Furthermore, the approach disambiguates w.r.t. similarity of contexts, prominence of entities and context windows. Unfortunately, the approaches presented so far are either not efficient enough (i.e. runtime lacks [8]) to handle web-scale data or do not deliver the expected extraction quality based on specific

[2] `http://lod-cloud.net/state/`

[3] `http://www.mpi-inf.mpg.de/yago-naga/yago/`

Linked Data sources [34]. Recently, Cornolti et al. [8] presented a framework for benchmarking NED approaches. The authors compared six existing approaches against five well-known datasets on different tasks and with different measures.

Information Extraction from *templated web-sites* is mainly related to the field of wrapper induction. Early approaches to learning web wrappers were mostly supervised (e.g., [18,11]). Recently, Crescenzi et al [9] described a supervised framework that is able to profit from crowd-provided training data. The learning algorithm controls the cost of the crowdsourcing campaign w.r.t. quality of the output wrapper. However, these novel approaches miss the opportunities related to existence of Linked Data, and the semantic consistency of the extracted data is out of their scope of interest.

In order to accomplish the vision of the Semantic Web, Gentile et al. [12] presents an approach for learning web wrappers that exploit Linked Data as a training data source for their wrapper induction framework. However, the process they adopt consists of a variety of manual steps and is thus very time consuming.

(2) *Search Query Support.*Auer et al. [25] describe a method to enrich search queries via a conjunctive extension based on the underlying semantic ontology. This approach is able to retrieve entities and documents provided only with a description instead of a search query. This leads to results without an overlap of keywords between query and document.

Besides keyword-based search queries, some search engines also understand natural language questions. Question answering is more difficult than keyword-based searches since retrieval algorithms need to understand complex grammatical constructs. Unger et al. [33] present a manually curated, template-based approach to match a question against a specific SPARQL query. They combine natural language processing (NLP) capabilities with Linked Data which leads to good benchmark results w.r.t. the question answering on Linked Data benchmark (QALD)[4].

(3) *Information Retrieval/Hybrid Search.*Popular search engines like Google or Yahoo! have answered search requests based on keyword queries for a long time. For a retrospective of existing information retrieval methods the interested reader may refer to standard literature [21]. However, the development of Semantic Web technologies lead to search engines being more conversational than traditional keyword-based engines [1].

Apart from those document- and keyword-centric approaches, the Linked Data movement has developed diverse strategies to leverage the advantages of semantic knowledge. Based on the underlying semantic structure of Linked Data, He et al. [14] developed an approach that transforms search queries to semantic graphs and tries to match those against the Linked Data graphs of the underlying dataset.

Furthermore, `http://swoogle.umbc.edu` represents a first prototype of a semantic search engine. Ding et al. [10] described the different search strategies to find instances via, e.g., term, document or ontology searches. Since

[4] `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald`

this application was updated in 2007 for the last time and only consists of a comparably small corpus of documents and Linked Data, it cannot be considered as a web-scale approach.

`http://sindice.com/` [7] is a more recent approach that scans the Semantic Web in order to build a semantic web index that is searchable and queryable via SPARQL. Unfortunately, the underlying database does not comprise full-text information and thus cannot answer a broad range of queries.

(4) *Ranking*

The procedures and algorithms described before are capable of delivering an unordered set of search results to the user. However, the increasing number of documents available on the Web leads to a tremendous growth of search result sets. Following Smyth et al. [31], most users tend to look only at the first few results. To aid finding relevant information within the first few places, ranking algorithms need to be deployed.

Well-known representatives for Web document ranking algorithms are the Hypertext-Induced Topic Search (HITS) algorithm [20] and PageRank [5]. Both calculate the relevance of a search result based on the Web link graph and are also very scalable algorithms.

Already in 2002, Mayfield et al. [23,29] described a first approach combining information retrieval with semantic inference mechanism. Furthermore, they present an algorithm which ranks Semantic Web entities with regard to trust information.

Moreover, an extension to the PageRank algorithm using Linked Data knowledge has been described by Julia Stoyanovich [32]. Extracting semantic knowledge from a Web document and combining this with an underlying ontology has shown to improve ranking quality. Unfortunately, this version of the algorithm is not able to scale on Web data.

Furthermore, ReConRank [17] is a highly efficient algorithm based on the PageRank algorithm. It considers provenance information while ranking, leading to more trustworthy result lists. This algorithm is based on semantic sub-graphs whose size influences efficiency and precision of results.

The ranking algorithms described so far are independent of the underlying query which can steer those towards a loss of information. Gupta et al. [27] introduced an approach that enriches the query based on Linked Data in order to find, e.g., polysemes and synonyms. Afterwards, the ranking works on a context-ordered index retrieving an initial sorting of the documents, which are finally sorted according to their similarity to the query.

Moreover, xhRank [13] proves that a combination of semantic information from a Linked Data graph can lead to an improved ranking. The position, morphological features and structure of an entity within a query are used to reorder certain documents from the search result list.

Past attempts combining Linked Data and information retrieval techniques suffer from either performance leaks and high quality results with respect to Web-scale datasets or a missing holistic concept that is able to bring search technology to the next level.

3 Problem Statement and Contributions

The aim of this doctoral work is an information system/search engine framework that will address the following working domains:

(1) Initially, the proposed system needs to link crawled Web data with Semantic Web knowledge. This task can be performed by NER, NED and RE algorithms. Therefore, two types of Web pages need be distinguished: templated sites like actor pages from `http://www.imdb.com/` and unstructured Web pages like news articles from `http://www.nytimes.com/`. In this thesis, two *Information Extraction* approaches have been developed, which are described in Section 4.

(2) After the data is provided, the user has to be enabled to search it. An effective way to do so is to provide the user with a input field-like interface they are used to. As the user begins typing into the search input field the framework should present different search query suggestions. This *auto-completion* does not only speed up searching but also teaches the user which kind of queries the search engine framework understands. Moreover, this can lead to a reeducation of users' search behavior from short keyword-based searches to longer natural language queries or even real search questions. An auto-completion approach which *supports the query generation* will be developed in the next stage of the PhD work using linked knowledge.

(3) The search functionality to be developed in this thesis is going to be *hybrid*, i.e., simultaneously performing a full-text,e.g., Lucene-based[5], and an entity search. Different entity search algorithms need to be developed based on the significantly different data structures and problems arising from them. While full-text search is a well-studied field, as shown in Section 2, entity search on Linked Data has only been in the focus of research for about 10 years. A hybrid search engine is currently under development and will be evaluated against the recently published QALD-4 benchmark.

(4) Finally, when appropriate Web pages and Semantic Web entities have been found, the user wants them to be presented according to their relevance. *Ranking* algorithms aim to reorder result list with respect to one or more sorting criteria. Scientifically sound methods for classical information retrieval are already present and the most important ones can be found in Section 2. However, principles creating a combined ranking of full-text and semantic search results need to be investigated within this doctoral thesis. Therefore, we aim at creating an machine learning-based interweaving of several well-known ranking algorithms.

Combining the advantages of information retrieval methods and Linked Data technologies will overcome the information flood problem. The union of highly scalable retrieval algorithms and effective rankings is able to increase the users search experience. A formalisation of the approach is currently in progress.

[5] `http://lucene.apache.org/core/`

4 Research Approach and Initial Results

Central to this PhD work is to answer *how a search engine can benefit from the Linked Data paradigm?* Diverse technologies like RDFa, micro-data and HTML5 semantic annotations have been introduced to enrich Web data for a better user experience and machine interoperability. However, to the best of our knowledge there is no information retrieval architecture that uses the advantages of this technology holistically. Moreover, some search pipeline steps for the Web of Data need to be revised in order to perform efficient and effective searches.

To meet this obstacle, the presented thesis introduces a pipeline architecture for a Linked Data-based search engine, as depicted in Figure 1.

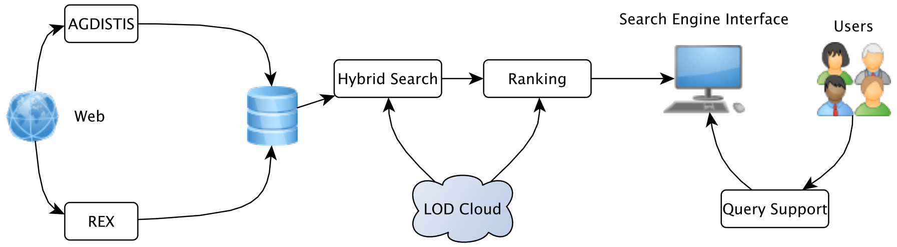

Fig. 1. Overview of the proposed information system architecture.

The starting point of the proposed architecture is a two-fold data acquisition strategy based on a highly efficient, state-of-the-art industry Web crawler provided by our research partner *Unister GmbH.*

First, *unstructured Web pages* from the crawled dataset, e.g., provided texts from news portals or agencies, are annotated by a standard NER algorithm [19] followed by a novel NED approach AGDISTIS [34]. This NED approach has been developed to support arbitrary Linked Data knowledge bases to ensure future developments. Moreover, AGDISTIS uses several NLP techniques to identify a set of candidate entities and identifies the correct with the help of the graph-based HITS algorithm. To prove the quality of AGDISTIS' results several corpora have been generated, evaluated and published. These corpora, called N^3 [28], use the state-of-the-art serialization format *NIF* [15] following the "eating our own dogfood" paradigm inherent to the Semantic Web community. N^3 are expected to form a novel gold standard in the areas of semantic named entity recognition and disambiguation. Using N^3 and other well-known datasets, AGDISTIS has been proven to outperform the state-of-the-art algorithm AIDA [16] by up to 16% F-measure. In the future, AGDISTIS will be evaluated against the framework of Cornolti et al. [8] to provide a more comprehensive evaluation.

Second, *templated Web pages*, e.g., `http://www.imdb.com`, have been identified as another important source for answering user searches. Therefore, REX [6] has been developed during the early stage of this PhD work. It is a web-scale semantic relation extraction framework capable to identify known as well as novel relations on Web pages creating RDF out of them. REX combines a well-known wrapper induction technique [9] for extracting XPath expressions, AGDISTIS as its NED algorithm and a consistency checker for the extracted relations based on ad-hoc generated schemas. It has been shown that REX is able to generate new Linked Data triples with a precision of above 75% [6].

The resulting data from both pre-processing steps will serve as the underlying dataset for future research steps together with knowledge from the LOD Cloud.

Concerning the users' need for exploring the data space, the next step is to *support the formulation of queries*. A huge potential within classical search engines is contained in inexact search queries, e.g., in terms of given a description only or a question. Standard search engine methodologies fail at this point due to not being able to match keyword queries. In this thesis, we will support query formulation by providing on-the-fly recommended queries based on the real-time user input. It is planned to use Linked Data such as *BabelNet*[6] to find polysemes and synonyms within a query and thus enhancing the understanding of what the users actually mean. Furthermore, three different standard approaches as well as a Linked Data-based grammar will be compared and evaluated against each other. Another by-product of an according auto-completion approach is to teach the user which queries a search engine understands.

The research field of information retrieval/search and ranking has so far only been analysed theoretically within this doctoral work. In this thesis, a hybrid search engine is going to be implemented, i.e., an engine comprising a full-text information retrieval system enhanced by extracted Linked Data and a stake of LOD Cloud-based entity search. Especially, the keyword-based search engine *SINA* [30] will be a starting point for further research.

With respect to ranking algorithms, this PhD work focuses on two different research plans. At first, a semantic extension of graph-based authority calculating algorithms will be investigated. Therefore, a master thesis has been looked after which analysed a context-driven enhancement of Stoyanovich's work [32]. Initial results show an improvement compared to the baseline using the plain PageRank algorithm. In parallel, an ensemble learning approach of Semantic Web-based ranking algorithms will be evaluated.

To summarize, the aforementioned steps will help building an integrated information system leveraging search engine performance using Linked Data. Additionally–due to strong industry needs–this framework is going to be used in a real-life environment with web-scale amounts of users. Finally, most of the source code will be published as open source and can be downloaded via the projects homepage[7].

[6] http://babelnet.org/

[7] http://aksw.org/RicardoUsbeck

5 Evaluation Plan and Conclusion

This PhD work is dimensioned for three years. After intense literature reviews in the beginning of the first year the need for annotated Web data has been identified. As a logical consequence, the development of AGDISTIS and REX had been finished by the end of the first year. Alongside, a gold standard (N^3) has been created to be able to evaluate the approaches mentioned above.

The second year will be used for developing and assessing the corresponding search and ranking procedures. To measure the quality of the *auto-completion* technology, we assess different real-world query logs from our industry partner. Thereby, we analyze how much characters are need to understand the query correct. Additionally, we focus on the efficiency of the system in terms of milliseconds to react on a pressed key.

Considering the ranking evaluation, we will use standard precision, recall and f-measures as well as rank comparision measures, e.g., mean reciprocal rank. The underlying data is provided by the industry partner through human rater assessments and several comparisons to real-life search engines, e.g., Google or Wolfram Alpha.

Afterwards, the combined pipeline itself will be evaluated in a qualitative study using professionals and end users. Therefore, empirical methods like Likert-scale questionnaires and direct relevance feedback will be used.

Next to refining already submitted work and optimizing the source code to meet industrial production standards, the developed approaches and algorithms will be refined in a spiral way if unpredictable results occur. Thereby, upcoming ideas will be interweaved with the presented schedule creating a closed loop consisting of research question, development, evaluation and new research questions.

Acknowledgements

Gefördert aus Mitteln der Europäischen Union

Europa fördert Sachsen.
ESF
Europäischer Sozialfonds

This work has been supported by the ESF and the Free State of Saxony.

References

1. Keynote at google i/o 2013.
2. SPARQL query language for RDF. Technical report, World Wide Web Consortium, January 2008.
3. Ben Adida and Mark Birbeck. RDFa primer 1.0 embedding RDF in XHTML. W3c working draft, W3C, October 2007.
4. Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N. Mendes, Bert Van Nuffelen, Claus Stadler, Sebastian Tramp, and Hugh Williams. Managing the life-cycle of linked data with the lod2 stack. In *International Semantic Web Conference (2)*, pages 1–16, 2012.

5. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *COMPUTER NETWORKS AND ISDN SYSTEMS*, pages 107–117. Elsevier Science Publishers B. V., 1998.
6. Bühmann, Usbeck, Ngonga Ngomo, Saleem, Crescenzi, Merialdo, Qui, and Both. Rex - web-scale extension of rdf knowledge bases. In *Submitted to 11th Extended Semantic Web Conference, May 25th, 2014 to May 29th, 2014 in Anissaras, Crete, Greece*, 2014.
7. Stéphane Campinas, Diego Ceccarelli, Thomas E Perry, Renaud Delbru, Krisztian Balog, and Giovanni Tummarello. The sindice-2011 dataset for entity-oriented search in the web of data. In *1st Int. Workshop on Entity-Oriented Search (EOS)*, pages 26–32, 2011.
8. Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 249–260, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
9. Valter Crescenzi, Paolo Merialdo, and Disheng Qiu. A framework for learning web wrappers from the crowd. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 261–272, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
10. Li Ding, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng, and Pranam Kolari. Finding and ranking knowledge on the semantic web. In *In Proceedings of the 4th International Semantic Web Conference*, page 156—170, 2005.
11. Sergio Flesca, Giuseppe Manco, Elio Masciari, Eugenio Rende, and Andrea Tagarelli. Web wrapper induction: a brief survey. *AI Communications*, 17(2):57–61, 2004.
12. Anna Lisa Gentile, Ziqi Zhang, Isabelle Augenstein, and Fabio Ciravegna. Unsupervised wrapper induction using linked data. In *Proceedings of the seventh international conference on Knowledge capture*, K-CAP '13, pages 41–48, New York, NY, USA, 2013. ACM.
13. Xin He and Mark Baker. xhrank: Ranking entities on the semantic web. In *ISWC Posters & Demos'10*.
14. Xin He and Mark Baker. A graph-based approach to indexing semantic web data. In *9th International Semantic Web conference (ISWC2010)*, 11 2010.
15. Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013*, pages 98–113. Springer, 2013.
16. Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, Edinburgh, Scotland*, pages 782–792, 2011.
17. Aidan Hogan, Andreas Harth, and Stefan Decker. Reconrank: A scalable ranking method for semantic web data with context. In *In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.
18. Andrew Hogue and David Karger. Thresher: automating the unwrapping of semantic content from the world wide web. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 86–95, New York, NY, USA, 2005. ACM.

19. Dan Klein and Christopher D. Manning. Fast exact inference with a factored model for natural language parsing. In *NIPS*, pages 3–10, 2002.
20. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 9 1999.
21. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. 2008.
22. Frank Manola and Eric Miller, editors. *RDF Primer*. W3C Recommendation. World Wide Web Consortium, February 2004.
23. J. Mayfield and T. Finnin. Information retrieval on the Semantic Web: Integrating inference and retrieval. In *Workshop on the Semantic Web at the 26th Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Toronto, Canada, 2003.
24. Pablo N. Mendes, Max Jakob, Andres Garcia-Silva, and Christian Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics)*, 2011.
25. Axel Ngonga. Generating conjunctive queries for keyword search on rdf data. In *In Sixth ACM WSDM (Web Search and Data Mining) Conference*, 2013. submitted.
26. Axel-Cyrille Ngonga Ngomo, Norman Heino, Klaus Lyko, René Speck, and Martin Kaltenböck. Scms–semantifying content management systems. In *The Semantic Web–ISWC 2011*, pages 189–204. Springer, 2011.
27. Dr. A.K.Sharma Parul Gupta. Ontology driven pre and post ranking based information retrieval in web search engines, 2012.
28. Michael Röder, Ricardo Usbeck, Daniel Gerber, Sebastian Hellmann, and Andreas Both. - a collection of datasets for named entity recognition and disambiguation in the nlp interchange formatN^3. In *LREC*. European Language Resources Association (ELRA), 2014.
29. Urvi Shah, Tim Finin, Anupam Joshi, R. Scott Cost, and James Matfield. Information retrieval on the semantic web. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, 2002.
30. Saeedeh Shekarpour, Edgard Marx, Axel-Cyrille Ngonga Ngomo, and Sören Auer. Sina: Semantic interpretation of user queries for question answering on interlinked data. *Submitted to Journal of Web Semantics*, 2013.
31. Barry Smyth, Evelyn Balfe, Oisin Boydell, Keith Bradley, Peter Briggs, Maurice Coyle, and Jill Freyne. A live-user evaluation of collaborative web search. In *In IJCAI*, pages 1419–1424, 2005.
32. Julia Stoyanovich, Srikanta J. Bedathur, Klaus Berberich, and Gerhard Weikum. Entityauthority: Semantically enriched graph-based authority propagation. In *WebDB*, 2007.
33. Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. Template-based question answering over rdf data. In *Proceedings of the 21st international conference on World Wide Web*, pages 639–648. ACM, 2012.
34. Usbeck, Ngonga Ngomo, Roeder, Auer, Gerber, and Both. Agdistis - agnostic disambiguation of named entities using linked open data. In *Submitted to 11th Extended Semantic Web Conference, May 25th, 2014 to May 29th, 2014 in Anissaras, Crete, Greece*, 2013.