STATISTICAL EXTRACTION OF MULTILINGUAL NATURAL LANGUAGE PATTERNS FOR RDF PREDICATES: ALGORITHMS AND APPLICATIONS

Der Fakultät für Mathematik und Informatik
der Universität Leipzig eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet Informatik vorgelegt von

M.Sc Daniel Gerber

geboren am 16.12.1985 in Bad Schlema, Deutschland

Leipzig, den 04. August 2015

[04. August 2015 at 18:04 – classicthesis version 4.1]

BIBLIOGRAPHIC DATA

TITLE:
Statistical Extraction of Multilingual Natural Language Patterns for RDF Predicates: Algorithms and Applications

AUTHOR:
Daniel Gerber

STATISTICAL INFORMATION:
143 pages, 29 Figures, 29 tables, 10 listings, 1 appendix, 157 literature references

SUPERVISORS:
Prof. Dr. Ing. habil. Klaus-Peter Fähnrich
Dr. Axel-Cyrille Ngonga Ngomo

INSTITUTION:
Universität Leipzig, Fakultät für Mathematik und Informatik

TIME FRAME:
January 2011 - July 2014

ABSTRACT

The Data Web has undergone a tremendous growth period. It currently consists of more then 3300 publicly available knowledge bases describing millions of resources from various domains, such as life sciences, government or geography, with over 89 billion facts. In the same way, the Document Web grew to the state where approximately 4.55 billion websites exist, 300 million photos are uploaded on Facebook as well as 3.5 billion Google searches are performed on average every day. However, there is a gap between the Document Web and the Data Web, since for example knowledge bases available on the Data Web are most commonly extracted from structured or semi-structured sources, but the majority of information available on the Web is contained in unstructured sources such as news articles, blog post, photos, forum discussions, etc. As a result, data on the Data Web not only misses a significant fragment of information but also suffers from a lack of actuality since typical extraction methods are time-consuming and can only be carried out periodically. Furthermore, provenance information is rarely taken into consideration and therefore gets lost in the transformation process. In addition, users are accustomed to entering keyword queries to satisfy their information needs. With the availability of machine-readable knowledge bases, lay users could be empowered to issue more specific questions and get more precise answers.

In this thesis, we address the problem of Relation Extraction, one of the key challenges pertaining to closing the gap between the Document Web and the Data Web by four means. First, we present a distant supervision approach that allows finding multilingual natural language representations of formal relations already contained in the Data Web. We use these natural language representations to find sentences on the Document Web that contain unseen instances of this relation between two entities. Second, we address the problem of data actuality by presenting a real-time data stream RDF extraction framework and utilize this framework to extract RDF from RSS news feeds. Third, we present a novel fact validation algorithm, based on natural language representations, able to not only verify or falsify a given triple, but also to find trustworthy sources for it on the Web and estimating a time scope in which the triple holds true. The features used by this algorithm to determine if a website is indeed trustworthy are used as provenance information and therewith help to create metadata for facts in the Data Web. Finally, we present a question answering system that uses the natural language representations to map natural language question to formal SPARQL queries, allowing

lay users to make use of the large amounts of data available on the Data Web to satisfy their information need.

ZUSAMMENFASSUNG

Das Data Web hat eine enorme Wachstumsphase erlebt. Es besteht aktuell aus mehr als 3300 öffentlich zugänglichen Wissensbasen, die Millionen Ressourcen von unterschiedlichen Domänen, wie etwa Biowissenschaften, Verwaltung und Geografie, mit über 89 Milliarden Fakten beschreiben. In gleicher Weise wuchs das Document Web zu dem Zustand in dem ungefähr 4,55 Milliarden Webseiten existieren und im Tagesdurchschnitt 300 Millionen Fotos auf Facebook hochgeladen und 3,5 Milliarden Google Suchanfragen durchgeführt werden. Trotzdem existiert eine Diskrepanz zwischen dem Document Web und dem Data Web, weil zum Beispiel im Data Web verfügbare Wissensbasen im Regelfall nur von strukturierten beziehungsweise teilweise strukturierten Datenquellen extrahiert worden sind. Allerdings befindet sich der Großteil der Daten im Web in unstrukturierten Datenquellen, wie etwa in Nachrichtenartikeln, Blogs, Fotos, Forendiskussionen, etc. Als ein Resultat dieser Diskrepanz fehlt den Daten im Data Web nicht nur der Großteil der verfügbaren Informationen, sondern lassen Aktualität vermissen, da typische Extraktionsmethoden zeitaufwendig sind und deshalb nur periodisch ausgeführt werden können. Des Weiteren werden Provenienzinformationen nur selten berücksichtigt und gehen damit im Transformationsprozess verloren. Außerdem sind Nutzer an Schlüsselwort-Anfragen gewöhnt, um ihr Informationsbedürfnis zu befriedigen. Mit der Verfügbarkeit von maschinenlesbaren Wissensbasen werden auch unerfahrene Nutzer in die Lage versetzt, spezifischere Fragen zu stellen und genauere Antworten zu erhalten.

In dieser Arbeit beschäftigen wir uns mit dem Problem der Relationsextraktion, eine der wichtigsten Herausforderungen, um die Lücke zwischen Document Web und Data Web zu schließen. Dazu stellen wir vier Methoden vor. Erstens zeigen wir einen Distant Supervision Ansatz, der es erlaubt multilinguale natürlichsprachliche Repräsentationen von formalen Relationen zu ermitteln, die bereits im Data Web enthalten sind. Wir nutzen diese natürlichsprachlichen Repräsentationen, um Sätze im Document Web zu finden, die unbekannte Instanzen dieser Relation zwischen zwei Entitäten enthalten. Zweitens beschäftigen wir uns mit dem Problem der Datenaktualität, indem wir ein Echtzeit-RDF-Extraktionsframework für Datenströme vorstellen und dieses Framework anwenden, um RDF aus RSS Nachrichten-Feeds zu extrahieren. Drittens präsentieren wir ein neuartiges Fact Validation Verfahren, basierend auf natürlichsprachlichen Repräsentationen formaler Relationen, das nicht nur in der Lage ist, ein gegebenes Tripel zu verifizieren beziehungsweise zu widerlegen, sondern auch vertrauenswürdige Quellen dafür im Web

findet und zusätzlich ein Zeitintervall bestimmt, in dem das Triple wahr ist. Die Merkmale, die von diesem Algorithmus genutzt werden, um zu bestimmen, ob eine Webseite vertrauenswürdig ist, werden als Provenienzinformationen genutzt und helfen somit Metadaten für Fakten im Data Web zu generieren. Zum Abschluss präsentieren wir ein Question Answering System, das natürlichsprachliche Repräsentationen nutzt, um natürlichsprachliche Fragen auf formale SPARQL-Anfragen abzubilden und es damit unerfahrenen Nutzern ermöglicht, die riesigen Datenvolumen im Data Web nutzbar zu machen um deren Informationsbedürfnis zu befriedigen.

PUBLICATIONS

This thesis is based on the following publications and proceedings. References to the appropriate publications are included at the respective chapters and sections.

AWARDS AND NOTABLE MENTIONS

- **Best Research Paper Award** at ISWC 2014 for *AGDISTIS - Graph-Based Disambiguation of Named Entities using Linked Data*.
- **Best Student Paper Award** at ESWC 2014 for *Hybrid Acquisition of Temporal Scopes for RDF Data*.
- **Spotlight Paper** at ISWC 2012 for *DeFacto - Deep Fact Validation*.

CONFERENCES, PEER-REVIEWED

- 18th International Conference on Knowledge Engineering and Knowledge Management: "*Extracting Multilingual Natural Language Patterns for RDF Predicates*" [**?**]
- 11th International Semantic Web Conference, 2012: "*DeFacto - Deep Fact Validation*" [**?**]
- 11th International Semantic Web Conference, 2012: "*DEQA: Deep Web Extraction for Question Answering*" [**?**]
- 20th World Wide Web Conference, 2012: "*Template-based question answering over RDF data*" [**?**]
- 12th International Semantic Web Conference, 2013: "*Real-time RDF extraction from unstructured data streams*" [**?**]
- 4th Conference on Knowledge Engineering and Semantic Web, 2013: "*TBSL Question Answering System Demo*" [**?**]
- 22nd World Wide Web Conference, 2013: "*Sorry, I don't speak SPARQL — Translating SPARQL Queries into Natural Language*" [**??**]
- 8th International Conference on Language Resources and Evaluation: "*N3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format*" [**?**]
- 11th Extended Semantic Web Conference, 2014: "*Hybrid Acquisition of Temporal Scopes for RDF Data*" [**?**]

- 13th International Semantic Web Conference, 2014: "*AGDISTIS - Agnostic Disambiguation of Named Entities Using Linked Open Data*" [**?**]

BOOK CHAPTERS, PEER-REVIEWED

- Chapter "*From RDF to Natural Language and Back*" in Towards the Multilingual Semantic Web, 2014 [**?**]

WORKSHOPS, PEER-REVIEWED

- 1st Workshop on Web Scale Knowledge Extraction at International Semantic Web Conference, 2011: "*Bootstrapping the Linked Data Web*" [**?**]

UNPUBLISHED PAPERS

- 8th International Conference on Language Resources and Evaluation: "*Mapping text to ontology with DBpedia Lemon and BOA*" [**?**]
- Journal of Web Semantics: "*DeFacto - Multilingual and Temporal Deep Fact Validation*" [**?**]

ACKNOWLEDGMENTS

I would like to thank all of my colleagues with whom I jointly wrote the papers and articles that led to this work: Axel Ngonga, Sebastian Hellmann, Lorenz Bühmann, Tomasso Soru, Jens Lehmann, Mohamed Morsey, Christina Unger, Philipp Cimiano, Konrad Höffner, Michael Röder, Ricardo Usbeck, Anisa Rula, Matteo Palmonari and Denis Lukovnikov. Special thanks goes to my direct supervisor Dr. Axel-Cyrille Ngonga Ngomo. He continuously supported me through my Ph. D. work, gave advice and recommendations for further research steps and improvements. I would like to thank Prof. Dr. Ing. habil. Klaus-Peter Fähnrich for his scientific experience with the efficient organization of the process of a Ph. D. thesis and Prof. Dr. Sören Auer who helped me to get a scholarship, which made this thesis possible. This thesis was funded by the Medienstiftung der Sparkasse Leipzig. I would like to thank Stephan Seeger for providing me with that opportunity. Also, I would like to thank Michael Martin who supervised my master thesis and guided my way to this dissertation. I would like to thank my close friend Robert Remus who challenged my ideas constantly and always provided valuable feedback. Finally I would like to thank my family, Swen, Sylvana, David, Irene and Rolf who have supported me in all respects over the period of this dissertation and beyond.

Thank You.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

Part I

KNOWLEDGE EXTRACTION FROM UNSTRUCTURE DATA

The first part

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

0.1 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 **?**, URI and HTTP in order to provide linkable, valuable content. Whether provided by a SPARQL **?** endpoint or embedded in a Web page via RDFa **?**, 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 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 **?** 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 0.2, the state of the art in the areas of information retrieval and Linked Data-based search and ranking algorithms is pre-

1 http://searchengineland.com/google-hummingbird-172816
2 http://lod-cloud.net/state/

sented. The problems tackled in this thesis and its contributions are described in Section 0.3. Section 0.4 presents the already available approaches *AGDISTIS* **?**, which is a named entity extraction framework for unstructured Web pages, and *REX* **?**, 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 0.5 concludes with an outlook on the future research agenda.

0.2 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 **?** 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 **?**. 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 **?**) to handle web-scale data or do not deliver the expected extraction quality based on specific Linked Data sources **?**. Recently, Cornolti et al. **?** 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., **??**). Recently, Crescenzi et al **?** 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

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

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. **?** 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. **?** 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. **?** 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 **?**. However, the development of Semantic Web technologies lead to search engines being more conversational than traditional keyword-based engines **?**.

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. **?** 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. **?** described the different search strategies to find instances via, e.g., term, document or ontology searches. Since this application was updated in 2007 for the last time and only consists of a comparably small

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

corpus of documents and Linked Data, it cannot be considered as a web-scale approach.

http://sindice.com/ **?** 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. **?**, 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 **?** and PageRank **?**. 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. **??** 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 **?**. 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 **?** 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. **?** 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 **?** 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.

0.3 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 0.4.

(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 0.2, entity search on Linked Data has only been in the focus of research for about 10 years. A hybrid search engine is currently

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

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 0.2. 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.

0.4 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.

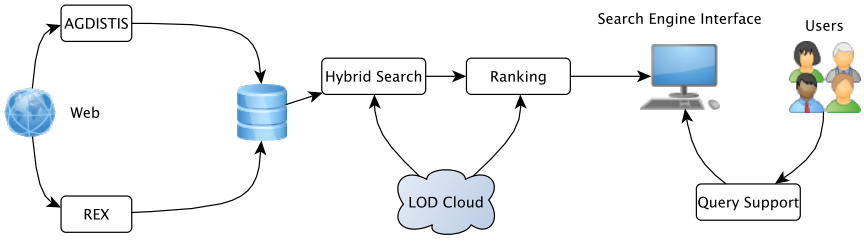

Figure 1: Overview of the proposed information system architecture.

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 **?** followed by a novel NED approach AGDISTIS **?**. 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 **?**, use the state-of-the-art serialization format *NIF* **?** 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 **?** by up to 16% F-measure. In the future, AGDISTIS will be evaluated against the framework of Cornolti et al. **?** 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 **?** 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 **?** 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% **?**.

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-

6 `http://babelnet.org/`

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* **?** 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 **?**. 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].

0.5 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.

7 http://aksw.org/RicardoUsbeck

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

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

BIBLIOGRAPHY

Part II

APPLICATIONS OF MULTILINGUAL NATURAL LANGUAGE PATTERNS

The second part of this thesis focuses on applications utilizing the multilingual natural language patterns generated by BOA and RdfLiveNews. This part therefore provides solutions for the quality and provenance problems of the current state of the Semantic Web. We first introduce DeFacto, a system which is able to verify or falsify a RDF triple. DeFacto does not only search for textual occurrences of parts of the statement, but also seeks to find webpages which contain the actual statement phrased in natural language. We then prove that our approach can be applied to multiple languages (English, French and German). Furthermore, we introduce a method to temporally scope facts. Finally, we demonstrate that these patterns can be used successfully in an ontology linking problem area in a question answering and a dictionary population task.

Part III

APPENDIX

BIBLIOGRAPHY

DECLARATION

This thesis is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

Universität Leipzig, August 2015

Daniel Gerber