Review of the Dissertation Thesis
by
Ivo Lašek
on
Information Extraction and Data Acquisition for News Articles
Filtering and Recommendation

Automated knowledge extraction has been the “Holy Grail” of data mining for decades, with direct, practical applications in information retrieval, recommender systems and content filtering. Intensive research over the past few years has resulted in approaches and algorithms suitable even for large scale, production quality systems. Traditional limitations of comparing surface forms of plain text documents have been supplemented by advanced semantics, with two main types of approaches: those based on distributional semantics and direct word co-occurrence, and those based on explicit knowledge bases, taxonomies and logical inference. This thesis deals with the latter, linking unstructured texts and structured knowledge bases, exploring practical limitations and applicability of such “linked” systems.

In this perspective, I. Lašek’s research on data acquisition, information extraction and entity linking clearly lies at the heart of the current research.

The dissertation is composed of seven chapters. Chapter 1 contains the problem statement and thesis objectives.

Chapter 2 introduces basic concepts behind entity recognition and linking and gives motivation for their use in practical systems. A case is made that linking relevant entities in articles (places, people, organizations) to a knowledge base improves overall system quality and content recommendations. The author successfully demonstrates this on two use cases: filtering news article streams, based on what entities were previously tagged as “interesting” by users. The other use case is enriching the LinkedTV service with background information from entities automatically detected and linked to web and video content.

**Question:** On page 11, it is mentioned a “text body” is extracted from HTML pages. This is a common source of noise when done automatically; how was the extraction of body and comments done?
Chapter 3, called “Background and State-of-the-Art”, covers a broad range of topics, from evaluation measures to MapReduce, existing implementations for web crawlers and entity linkers, RDF storage databases, to algorithms for supervised, semi-supervised and unsupervised entity linking. Although the breadth of the introduced concepts is quite large, the chapter is clearly written and well referenced, demonstrating a high level of competence acquired by I. Lašek in this domain.

Chapter 4 deals with data acquisition: crawling, transformation and data storage. It documents common pitfalls and struggles of using existing open source crawlers and Hadoop for parallelization. The chapter is technical, focused on engineering. Its ultimate contribution is an open sourced Java application, running as a public web service, able to crawl arbitrary LinkedData resources using a RESTful API.

**Question:** How much of this contribution was actually done by the author, as the headers in source code files mention “Ondrej Klimpera”, and none mention Ivo Lašek?

**Question:** The SVN code repository is hosted on the author's faculty page. Wouldn't a more standard hosting solution, such as GitHub, give more visibility and impact to these results?

Chapter 5 explores Wikipedia as the knowledge base specifically. The chapter covers topics of efficient indexing, data cleaning and preprocessing. Section 5.6 appears to contain the core of the academic contribution of this thesis: a novel algorithm called “structural context representation” and its optimizations. Given a text containing surface forms of entities with multiple candidates, the algorithm disambiguates said candidates as a whole and selects a single most appropriate entity for each surface form. Author introduces a modified “max” algorithm that has acceptable runtime even for longer texts with many surface forms.

**Question:** In Section 5.4.6 (pages 63-64), author describes his move away from SQL in favour of “in-memory processing”. Whether data is stored in memory or on disk is largely orthogonal to what language is used to query it (SQL databases work in-memory too). Can he motivate this move better? Similarly, why should GNU sort take an hour on a fairly small (4.6GB) file, what caused this slow performance (page 64)?

Chapter 6 evaluates the presented entity linking algorithms and implementations on several evaluation sets, some of them standard, some created specifically by the author. I especially liked the section with graphs on statistics of Wikipedia data, analyzing the distribution of entities, which display I. Lašek's understanding of the messiness of real world data. As the developed system is largely language-independent, Chapter 6 also explores its multilingual performance on Dutch and German inputs. An interesting error analysis is given for evaluation on the TAC Entity Linking Track challenge (page 90). Author notes that many surface forms are, in fact, not ambiguous, and even when they are ambiguous, simply choosing the most popular candidate is a hard-to-beat baseline algorithm. For this reason, author constructs their own benchmark with highly ambiguous surface forms, and evaluates several algorithms this way. The algorithm proposed in this thesis is demonstrated to perform better than DBpedia Spotlight, a state of the art open source tool.

**Question:** How did the results achieved on the TAC Entity Linking challenge compare to other contestants?

**Question:** “Type determination” makes its first appearance on page 87, without being introduced or described. What is it, what exactly is being evaluated here?

Finally, chapter 7 summarizes main contributions of the thesis. It sketches exciting directions for future work, especially in the area of entity recognition (candidate generation). Entity recognition
was relegated to 3rd party tools in this thesis, but at the same time recognized to be critical for good overall linking performance.

**Question:** on page 101, two newly developed benchmarks are listed as part of the main contributions of the thesis. Are these benchmarks public for other researchers to use? Have other people used them?

**Overall, the thesis is well structured.** I particularly appreciated the summary sections at the end of each chapter, recapitulating the chapter’s main points and achievements.

**On the other hand, the mathematical apparatus used in the thesis appeared self-serving.** Where I found the text to be too ambiguous and looked for clarity in formulas, I generally failed. For example, the “\("\)" symbol for set difference is apparently overloaded to mean something else (what?), because it doesn’t make sense as used (formulas 5.2, 5.13, 6.1 and elsewhere, somehow mixing a set of entity mention triples with a set of single entities). The set of entities E is defined to contain all entities mentioned on Wikipedia (page 54), and then E₀ is defined as its subset of all entities mentioned on Wikipedia (eq. 5.14, page 59)... what is the difference between E and E₀?

What is the “50% probability” mentioned on page 83 actually a probability of? I appreciate the objectives of the thesis are exploratory, focused on system integration, evaluation and practical aspects, but in my opinion, the plain text of this manuscript would have been just as clear on its own, without the formulas and equations.

In summary, the thesis explores an exciting area of enriching plain text documents with information from structured knowledge bases, with good results and a strong publication record (eight peer-reviewed journal and conference papers). During the course of his research, the author and his colleagues also developed several software tools: a Linked Data crawler, a Wikipedia indexer, an entity linking tool SemiTags. These practical contributions meet the modern criteria for open research, with published source code and public web APIs. As a pleasant bonus, though not directly connected to the goals of this dissertation, I. Lašek started off a Czech version of the structured Wikipedia project, the Czech DBpedia.

**As a general conclusion, I consider Ivo Lašek to have clearly demonstrated his ability to work independently and creatively in his area of research.** His dissertation thesis meets the requirements of a successful research work. **I thus recommend I. Lašek’s thesis to be accepted** and wish him success in his follow up research career.

Prague, 4 Jan 2015

Radim Řehůrek
REVIEW OF THE DISSERTATION THESIS SUBMITTED BY IVO LAŠEK
INFORMATION EXTRACTION AND DATA ACQUISITION FOR NEWS ARTICLES
FILTERING AND RECOMMENDATION

This thesis is focused on the filtering of news articles and articles' recommendation according to extracted data and linked information.

The topic of the thesis is actual and belongs to the broader area of semantic search and knowledge extraction.

The structure of the thesis is well defined. Each chapter contains information needed in the following chapters. The main theoretical content is located in Chapters 4 and 5, with experimental evaluation in Chapter 6. The dissertation contains several mistakes and typos which decreases the formal quality of the dissertation. Moreover, the dissertation contains redefinitions of few terms and methods. Fortunately, the definitions are the same, but it seems that the dissertation has been modified in the last minute and these mistakes were not amended.

The objectives of the dissertation, as were defined, may be considered as complete. But the comparison of the methods for named entity recognition (objective 3) is very short and compares only two methods on Dutch and German samples. Both tested algorithms were developed mainly for English language and this comparison is odd, even with respect to the used training datasets. The second comparison – context aware disambiguation – is completely relevant and very interesting.

The methods used in the dissertation follows the standard methodology used in this area. Standard methods for entity linking and knowledge base building are used, but their implementation is detailed and many optimizations have been introduced. Moreover new principles based on the Map-Reduce paradigm are tested with good results.

The results presented in this dissertation may be divided into several groups. Some results are interesting but not new or revolutionary, i.e. mainly Wikipedia statistics and the comparison of the disambiguation state of the art tools on German and Dutch texts. In addition to that, the evaluation of the SemiTags tools on 10 German articles is small. The other experiments are very good, mainly the construction of the Highly Ambiguous Benchmark, as well as, the related evaluations. Few contributions of this dissertation are important and were already mentioned in the review. For example, the efficient approach for linked entity occurrences using in-memory computation and the creation of the Highly Ambiguous Benchmark.

The author of the dissertation proved his ability to write research papers of which one paper was published in a Journal indexed on Scopus, in addition to a few others that were published in...
indexed conferences. Unfortunately, part of the work presented in this dissertation is not relevant or beneficial to the field, but few parts are very interesting and may be evolved in further research.

The author of the dissertation proved the ability to conduct research in the area of knowledge extraction and to achieve scientific results which were published in journal and conference proceedings. Despite minor problems with formal state of this dissertation, mainly English typos, errors and a few irrelevant parts, I evaluate the dissertation as acceptable.

In accordance with par. 47, letter (4) of the Law Nr. 111/1998 (The Higher Education Act) I do recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

Ostrava, 28.1.2015

doc. Ing. Jan Platoš, Ph.D.
Department of Computer Science, FEECS
VSB-Technical University of Ostrava
Review of Dissertation by Ivo Lašek

Information Extraction and Data Acquisition for News Articles Filtering and Recommendation

Up-to-datedness of the dissertation.

At the time of writing, the thesis was up-to-date and competitive with existing approaches. Most of the published articles of the author are from 2012 and show good progress over the state-of-the-art. Some developments such as the Workshop Series on NLP & DBpedia are highly pertinent, but definitely too recent to be mentioned in the thesis. Thus, I don't hold it against the thesis author. One minor remark, that I have, is about the missing citation of Kontokostas et. al. 2012, “Internationalization of Linked Data: The case of the Greek DBpedia edition” in Chapter 4 regarding the Czech DBpedia. This could replace the link to the Wikipage by C. Sahnwaldt and would be more appropriate.

Formal structure and organization of the dissertation.

The outline of the thesis is straightforward and appropriate.

Regarding Chapter 3, I have to admit that I was quite surprised to see a Chapter about Background, State of the Art AND Related Work at the same time. Although this is quite unusual for my understanding, the author manages quite well to combine all three things. Thus, the chapter helps to give a good overview of the work in the thesis, as well as put its contribution into the right scientific context.

Chapter 5 contains a formalization of the main problem as well as a description of algorithms. While the formalization is excellent and captures the problem adequately, the different algorithms are presented in a confusing way. In my opinion, the presentation would have been clearer with a section which describes all indexes used, followed by a section with all algorithms used. At the moment these descriptions are sequentially intertwined and document more the evolution and history in the order the work has been done instead of focusing on the result.

In Chapter 6, Figure 6.1, 6.2 and 6.3 b) contain log numbers at the y-Axis instead of the real values, which is quite unpractical. I suggest to use 100 instead of 2. Overall, I was wondering about the descriptiveness of the presented tables (e.g. 6.1) having excerpts for the first five or ten entries. Sometimes they add up to 100%, sometimes not. Furthermore, I am uncertain, what we can learn from these tables as it often remains unclear what the exact total numbers are, and which statistics they were created from. A clear reference to Chapter 3 is missing.

As a minor comment, the reader has to look up CONLL-2003 to get to know the languages (English and German) of the dataset and it was difficult to realize that the NER used on Dutch text was trained on German data.
Completion of the dissertation objectives.

The objectives of the dissertation are given in Section 1.3. Given the careful phrasing of the contributions, the objectives are all fulfilled. Clear performance indicators are not given and thus the objectives remain vague overall. From my personal experience, I would judge the challenge of the set objectives as sufficient for the level of a dissertation.

Assessment of the methods used in the dissertation.

Chapter 4, 5 and 6 describe the main contribution of the work.

Chapter 4 contains a selection of methods to assemble a crawl of Semantic Web data. The resulting performance measure stay quite superficial. Only two data sources are crawled and then speed is compared. Of course, the implementation has to honor the delay given in the robots.txt. However, the assessment could have been done with this delay factored out to achieve a better comparison between dump and individual URI download.

Chapter 5 represents the core work of the thesis and describes in detail the data structures and algorithms. The first part describes in which way the data basis for the disambiguation models are gathered. The transition from table 5.1 to 5.4 is quite vague and it remains unclear what data is cleaned in which way. Normally, redirects can be resolved to "real" pages instead of just being removed. So in the CSV files all occurrences of redirects could have been replaced with the target article link. Looking at Figure 5.7 I could not understand exactly what the SQL query does. Overall, the 85% reduction can not be traced by reading the description and is unclear (1) how 9,709,063 entities were boiled down to 1,565,389 and why the article count of 4.4 million could not be utilized to index 4.4 entities. [91] add a cleaned version of the title to their index, which would have been an easy optimization. A total number of paragraphs and surface forms is missing in Table 5.4.

Chapter 6 starts with statistics derived from Wikipedia, especially entity per paragraph and also surface form to entity distribution as an ambiguity measure. Looking at the number, there are some discrepancies, which I was unable to resolve. While the text on page 81 talks about 3,734,848 normalized surface forms, Table 6.2 lists 3,697,357 and I am unable to judge which ones are correct. Overall normalization merges around 600,000 forms which is around 0.85% of 70 million entity mentions and not 0.71% as mentioned on page 81. For Table 6.2 the surface form "2006" is given as the most ambiguous surface form, but this is not reflected in the table, as the number add up to 100%. Shouldn't this be 99%?

In the summary of Chapter 6 conclusions are drawn, which do not directly follow from the data. From a logical perspective it is completely clear that proper candidate generation is a prerequisite of correct disambiguation as a sine qua non condition, i.e. it is impossible to assign a correct link, if it isn’t within the candidate set. However, the measurements presented in Figure 6.15 just give a hint in this direction. The 37.9% of the piechart remain vague as it is left open, how many of the 11 queries (i.e. 11 of 29 total errors) failed because of the deviant surface form and how many failed because the entity was not in the KB.
This is especially important since only 1.6 million entities are left from the cleansing of the index (potentially around 4.4 million articles in Wikipedia).

**Evaluation of the results and contributions of the dissertation.**

Chapter 3 gives a well structured overview, which allows even inexperienced reader to understand the domain.

Chapter 4 contains a description of the creation of the Czech DBpedia, which has been done with great skill and has created a high-impact resource by providing Czech linked data. Also the crawling algorithms are described.

Chapter 5 lists a good overview of the indexes and the algorithms used in the thesis. As the algorithms are well described and the tools have been provided as open-source, they are easy to reuse and are a significant contribution of the dissertation.

Chapter 6 analyses the index and algorithms given in Chapter 5. The description is sufficient to reproduce the measurements. Overall, I would have hoped for a more systematic approach. Especially, the influence of the candidate generation could have been revisited. A clear influence is obvious, but has not been measured. The trade-off between a clean, high-precision index versus a more broader not so clean index is not shown for example. This is left open for the next researcher that can draw upon this dissertation as a preliminary work.

**Remarks, objections, notes, and questions for the defense.**

**Remarks:**

Most of the errors noted above are potentially easy to correct and should be corrected before the thesis and the final submission.

**Questions for the defense:**

Q1: What was the motivation to choose max(score)/sum of scores as the confidence measure? What is your experience with this measure?

Q2: What was the motivation behind choosing the used measures for soft and 3d benchmarks? What do you think are the use cases for such benchmarks?

Q3: Regarding lessons learned: can you give clear guidelines when to choose which algorithm for disambiguation?
The overall evaluation of the dissertation.

The thesis has been written at an early stage of entity linking, where the problem as such was not well understood or formalized. While the thesis does not produce a significant progress of the state of the art in this area, it offers a good insight into the engineering problems that are necessary to achieve the creation of such a system. All of the work done is described in sufficient detail for reproduction and the source code has been made available. Therefore the contribution is clear.

A major point, I have to criticize are the four disconnected parts Czech DBpedia, Semantic Web crawling, multilingualism and entity linking which are not well bridged. In principle, I am favorable towards theses with mixed and not aligned contributions, however, in this case more effort could have been put into the attempt to build a golden roof connecting all topics, e.g. using Wikipedia/DBpedia as a thread more prominently.

Overall, the thesis documents the creation of several practical Semantic Web systems and evaluates them, which makes it a valuable contribution.

Conclusion

The author of the dissertation proved the ability to conduct research and achieve scientific results. I do recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

Leipzig, den 4.03.2015

Dr.-Ing. Sebastian Hellmann

Head of AKSW/NLP2RDF Research Group at
University of Leipzig
Minor comments for individual sections:

2.1. Statistical approaches also hide internals from the developers, so any debugging is more opaque than with semantic filtering

2.1.2 could have cited the DBpedia Spotlight paper on this as the measure is similar.

2.2 should be written in past tense

3.1.3 Figure 3.1 is taken from a master thesis. This is good in general, but a bit unusual for the background chapter.

3.1.4.2 I am a bit puzzled how the Dot Product can be defined as the “intersection” of the vectors.

3.2.3 Sindice is missing, but mentioned in the next paragraph

3.4.5. “Pretty much state of the art” -> Is it or not?

Spelling:
content based -> content-based
non ambiguous -> non-ambiguous
6.4.0.1 should be 6.4.1
Figure 4.1 is tiny, please enlarge
Page 85: “incorrectly, even the” -> “incorrectly, also the”
sentence based -> sentence-based