

# Review of Jaroslav Kuchař's PhD Thesis "Rich Semantic Representations in Web Usage Mining"

---

## 1. Up-to-datedness of the dissertation

Web usage mining enables collection of actionable data based on observing interactions between users and content on the Web. Such data can be highly valuable for example in a commercial context. The core focus of the thesis consists in exploiting semantics in rich representations that link information about user interactions and descriptions of content items that users are interacting with. In this sense, the thesis goes beyond state-of-the-art approaches that do not take into account semantics of the objects and the irregular length of interactions performed by individual users. The dissertation presents novel approaches in this context using semantics for building and utilizing rich representations connecting users and content.

The dissertation covers theory, algorithms and evaluation as well as proof-of-concept implementations. The thesis contributions are centred around a methodology that covers a broad spectrum of relevant tasks, from data acquisition over semantic annotation and enhancement to utilization of rich semantic representations.

## 2. Formal structure and organization of the dissertation

The thesis is written in English, with language and grammar on a relatively good level. The dissertation is structured into four major chapters, organized in a logical way and has a smooth reading flow.

The first chapter contains introduction and motivation as well as a clear problem statement in terms of the main hypothesis supported by a set of research questions and a summary of the dissertation's contributions addressing those questions.

In the second chapter the author presents the background and state-of-the-art analysis. The background information is structured in a way that each subsection provides basic definitions of concepts used in the proposed solutions to each research question. The state-of-the-art analysis is also structured in accordance with the research questions and dissertation's contributions, and surveys related works in relevant areas.

Chapter three describes in detail the dissertation's contributions, where each section is dedicated to a specific problem and consists of theoretical considerations, proposed method, implementation and evaluation.

Finally, the last chapter contains the conclusion and directions for future work.

### 3. Completion of the dissertation objectives

The dissertation objectives are presented in terms of four research questions covering challenges in the four basic steps of the thesis' methodology: data acquisition, semantization and transformation, enhancement, and utilization. The contributions in the third chapter line up towards a comprehensive framework for the execution of each of those steps. The objectives set out for the thesis have been completed through a set of methods, algorithms, and experiments.

### 4. Assessment of the methods used in the dissertation

The methods used in the dissertation are scientifically sound. The thesis starts with a relevant motivation to the addressed topic, presents basic concepts, illustrates new ideas and approaches, and evaluates them against the state of the art. The evaluations are done, in most parts, empirically, through a set of well-defined experiments.

### 5. Evaluation of the results and contributions of the dissertation

The core contributions are presented in the third chapter. The evaluations of the proposed approaches are well described in the dissertation. A number of tables and plots were given to visualize and compare the results of the experiments with different methods. The extensive experiments carried out part of the thesis prove the ability of the author to master various evaluation techniques. The proposed approaches were published in several scientific articles, giving evidence that the author's work has found acceptance in the scientific community.

### 6. Remarks, objections, notes and questions for the defence

The overall impression is that the author has done a very good job on a technical level. The thesis has a good scientific baseline and shows the author's capability to carry out research in a systematic and methodical manner, and publish results, while at the same time showing authors' sound engineering skills.

Although examples are presented to illustrate various aspects of proposed methods and approaches, the thesis would have benefitted from the use of an integrated running example at the very beginning of the thesis, which could have been referenced from various sections. This way the connection between the various steps in the methodology would have been easier to follow.

The authors discuss to certain level of details various evaluations, however not very much is discussed about limitations of the proposed methods and approaches. Some details on limitations are given in the future work section, however it would be interesting to hear more about the limitations and their practical impact.

An interesting aspect to hear more about is also the practical implications of the work carried out in the thesis. For example, what kind of applications (and in which domains/verticals), if any, can be enabled with the results of the thesis, which were not possible before? How the

various algorithms/methods presented in the thesis could be integrated in a software package/library/system/framework in such a way that a typical programmer could make use of them on a regular basis to build relevant applications and services?

Some further specific questions:

- (p.36, 3.1.1. Definitions, Interest level)

What means negative interest? What was the reason for using [-1,1] numerical interval instead of using [0,1] scale?

- (p.37, 3.1.2 Data acquisition)

What is the difference between implicit and explicit user interactions (user feedback)? Is the type of interaction defined by the way of usage data collection?

- (p.61, 3.2.1 URI alignment)

Does the set of pre-defined SPARQL queries for the URI alignment have to be domain-specific? Who should compose these queries? Can there be any universal approach (e.g. with the help of searching by labels)?

- (3.2.1.3 Confidence values, page 64)

When computing a title confidence value, it is noted that the proposed metric can give false results because of comparing strings in different languages. Why cannot we specify language tag in pre-defined SPARQL queries and avoid this problem?

## 7. The overall evaluation of the dissertation

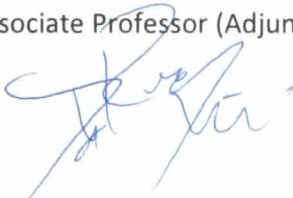
The author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 137/2016 (The Higher Education Act) I do recommend the thesis for the presentation and defense with the aim of receiving the Ph.D. degree.

November 2016

**Dumitru Roman, PhD**

Senior Research Scientist, SINTEF, Norway

Associate Professor (Adjunct), University of Oslo, Norway





Student: Jaroslav Kuchař

Title: Rich Semantic Representations in Web Usage Mining

With a growing number of users and the size of the web, the need of understanding their interaction becomes very important. Author challenges automation of semantic enrichment of information about users, content and connections. This is an important problem which is intensively studied worldwide.

Formally the 180 page thesis is organized in four chapters Introduction, Background and State-of-the-Art, Contributions and Conclusions. The main part (about 100 pages) is Chapter 3 Contributions. It is divided into 5 parts: 3.1 Acquisition of User Interactions, 3.2 Semantization and Propagation, 3.3 Graph-based Data Enhancement, 3.4 Personalized Selection of Entities and 3.5 Rule-based Preference Learning and Recommendations.

Problem statement consists of main hypothesis "Rich representations that link usage data and content descriptions profit from semantics, which enable to exploit extensive amount of linked information." This hypothesis is substantiated into four research questions (objectives):

- Can we incorporate semantics to processing and aggregating users' interactions and provide unified extended representation of relations between users and content?
- How can we extend already existing set of features describing the content by additional features using existing semantic sources and transform them to a semantic representation?
- Is there a possibility to enhance a semantic representation about new links, while taking into account semantics of links and temporal information?
- How can we utilize the rich semantic representation connecting users and content for selection of content items or recommendation according to users' preferences?

I think that objectives are well posted and research question describe overall methodology by which author shows their fulfillment.

Methods and technologies used are well suited for the challenging hypothesis. Main contribution consist of:

- a method and implementation for acquisition and aggregation of user interactions. Prototype was evaluated in domains of web analytics, Smart TVs and recommender systems.

- a mapping of domain specific content to a knowledge base using Linked Open Data cloud. Method was evaluated on a multilingual movie ratings dataset and published in a public dataset. It was evaluated within trials of the LinkedTV EU Project.
- a method for prediction of links and selection of the most relevant target within one dataset with the concept of forgetting factor to decrease the influence of older links. The link prediction algorithm was evaluated on Web APIs directory and a movie ratings datasets.
- to have self-explainable and justifiable user preferences and recommendations author uses rule learning. Evaluation was done on domains of Smart TV and News recommendation challenge.

All algorithms and data are publicly available and experiments are repeatable. Published results are in refereed sources, 4 of them cited in WoS and 10 in Scopus.

I would like to ask the candidate to comment more on experimentation methodology, e.g. on other methods (base-line, alternative methods) which are relevant to tasks and data used.

I would appreciate comments on content boosted matrix factorization which is also relevant to processing enriched data.

As the work is sound and clearly shows creative scientific development, I do recommend to admit the PhD thesis of Jaroslav Kuchař for presentation at the defense.

Prague November 29<sup>th</sup>, 2016



Peter Vojtáš

**PD Dr. Gerhard Wohlgenannt**  
University of Economics and Business  
Institute for Information Business  
1020 Vienna, Austria  
☎ +43 (676) 8213 5228  
✉ gerhard.wohlgenannt@wu.ac.at

**To: Czech Technical University in Prague**  
Faculty of Information Technology  
Prof. Pavel Tvrđík  
Thakurova 9, 160 00 Prague 6

November 15th, 2016

**Review of PhD Thesis  
submitted by  
Ing. Jaroslav Kuchař**

**Title: "Rich Semantic Representations in Web Usage Mining"**

## 1. Up-to-datedness of the dissertation

With the growth of the Web, and the increase of interaction between Web users and Web content, the topic of Web Usage Mining is obviously a relevant and very important one. Traditional systems rely for example on the analysis of Web server logs, but with modern interactive and multimedia Web content, new ways to generate usage data, and utilizing usage data, are necessary.

The thesis of Jaroslav Kuchař addresses different topics in the area of Web Usage Mining, building on rich semantic representations of usage data. Those rich representations connect users and Web content, and also enrich the content with further semantics.

## 2. Formal structure and organisation of the dissertation

The thesis is organized into 4 main chapters. In the *Introduction* chapter, the author gives the motivation, the problem statement and an overview of contributions of the thesis, as well as an outline the remainder.

Section 2 then gives a brief introduction of the theoretical background, and describes related work regarding the 5 contributions.

The main section is about the contributions of the thesis. The section is split into 5 contributions, which are largely independent and evaluated on different underlying data sets, but which also interact and build on each other. Those 5 main contributions are: (i) work on capturing and analysing the interactions between users and content, mostly in the context of interactions of users with a *smart TV* systems. For example, the author provides methods on how to estimate the users interest in a specific content based on his previous interactions, and compares various machine learning techniques to train the system. Contribution (ii) firstly presents an algorithm to link user ratings on films from a movie review dataset (extracted from Twitter data) to DBpedia. The author

uses SPARQL queries to find the movie in the knowledge base, and assigns confidence values to the linking information. In the second part of this contribution, the author discusses how to aggregate information from multiple semantic annotations / links to KBs for an *object*. The method propagates information in a taxonomy, and then merges that information in order to find likely class assignments for the object in question. Contribution (iii) focuses on the prediction of links within one dataset, esp. semantic datasets with have labelled (and multiple) links between nodes in the graph. In contrast to other work, time (or age of the link) plays a major role in the proposed algorithm. The author uses a tensor model and tensor factorization for link prediction. Contribution (iv) is about recommending and selecting resources within a graph of users and resources. In the proposed, again, method temporal information plays an important role – by using a configurable aging factor. Furthermore, the user defines preferences about link types (based on the semantic information of the link) which influences the recommendations. The maximum activation method (similar to spreading activation) is then used to select resources. Finally, in contribution (v) the author presents work in the field of preference learning algorithms and semantic-aware recommendations, using the rich semantic information generated in contributions (i)-(iv). Instead of black-box-based models, a system based on rule learning provides good results on recommendation tasks. The dissertation ends with a *Conclusions* chapter, which includes a summary of work, the contributions, and future work.

The thesis also includes a list of 161 references in the *Bibliography* section, and a list of reviewed publications of the author which are relevant for the thesis, and finally a list of other publications by the author. The list of 17 reviewed publications (co-)authored by the candidate underpins the scientific relevance of the his work.

The thesis has been written in English language, while not a problem regarding understandability, there are still a large number of typos and especially of grammatical mistakes in the thesis; therefore the reviewer recommends another proofreading.

### 3. Completion of the dissertation objectives

The dissertation objectives are described in the 5 research questions in chapter 1, which are all related to different aspects of Web usage mining. The goals were solved sufficiently, for any of the research questions novel work beyond the state-of-the-art is presented by the author. For all research questions, the author described new methods, and evaluated those methods on real-world data sets. Often, implementations of the methods are released as public open-source tools (on github).

### 4. Assessment of the methods used in the dissertation

As the thesis consists of 5 contributions with regards to 5 related research questions, a plethora of methods is being applied. In contribution (i) the author uses heuristically defined rules, and also a genetic algorithm with two different fitness functions. Furthermore, in the evaluations he also applies classical machine learning techniques such as SVN or KNN as baselines. Contribution (ii) applies SPARQL queries for URI alignment, and presents an algorithm for semantic propagation and aggregation. Contribution (iii) suggests a method based on tensor models and tensor factorization for the link prediction problem. In contribution (iv), the author of the thesis develops a novel selection method for personalized recommendations, building for example on work on spreading

activation networks and the Ford-Fulkerson algorithm. Finally, in contribution (v) rule learning methods are presented, which build for example on Association Rule Mining and decisions trees, or use those as baselines, respectively.

## 5. Evaluation of the results and contribution of the dissertation

The contributions have already been described in detail in the previous sections. For each of the contributions, the author evaluates the proposed methods with real-world datasets, for example a tourism agency dataset for contribution (i), a movie ratings dataset for contribution (ii), data about users, APIs and mashups from the ProgrammableWeb platform (contributions (iii) and (iv)) or datasets from recommendation challenges in contribution (v). The author performed extensive evaluations of the proposed methods, also using a number of baseline approaches to compare against. Some of the methods and evaluations are limited to certain domains, the author will work on generalization in future work.

## 6. Remarks, objections, notes and questions for the defence

Some minor remarks about typos and grammatical errors were made in section 2. They will contribute to improving the overall quality of the presentation and the readability of the thesis. Furthermore, the topic the *privacy* seems very important to this work, as it processes and stores a lot of user-specific and potentially sensitive personal data. Those issues are briefly discussed in the work, but as a question of the defence, I would like to hear the opinion of the candidate on how to ensure privacy in this context.

## 7. The overall evaluation of the dissertation

The author of the dissertation proved the ability to conduct research and achieve scientific results. In accordance with par. 47, letter (4) of the Law Nr. 111/198 (The Higher Education Act) I recommend the submitted Ph.D. Thesis by Jaroslav Kuchař for the presentation and defence with the aim of receiving the Ph.D. Degree.

Best regards (in Vienna, 15th November 2016),

  
PD Dr. Gerhard Wohlgenannt