Extracting recipe ingredients from cookbooks

by

Torsten Knauf

A thesis presented for the degree of
Master of Science

Research Group for Communication Systems
at
Faculty of Engineering
Christian-Albrechts-Universität zu Kiel
Germany

31.03.2017

Supervisor: Prof. Dr.-Ing.Norbert Luttenberger
Dr.-Ing. Jesper Zedlitz

Abstract

Always do this one last, when knowing the things to praise yourself for :P

Acknowledgements

If I don't profit from nice people in these thesis, I have done something horrible wrong. So try to remember most of them here at the end... :)

Contents

Chapter 1

Introduction

The introduction points out, what this thesis is about, why it is actual useful and finishes with an overview of the structure of this thesis.

1.1 What is this thesis about?

In general extracting automatically information from textual resources is a new field of research. In contrast to knowledge discovery in databases the information are unstructured, what makes it hard to automatically extract usefull information. For example the world wide web consists mainly of textual information, enriched with some meta data for formating, but without clues of its content. At least without clues, which a machine can process further. This restricts heavily the possibilities of new applications as described in (Berners-Lee et al. 2000).

For changing the lack of machine-readability, an ontology is required, as well as the usage of the ontology on the textual information (Berners-Lee et al. 2000). An ontology for recipes is already in place (Schema.org 2016). But to deploy a strict ontology on textual information manually, for example through tagging, is slow and error prone (Erdmann et al. 2000).

Text mining deals with the automatic extraction of information from unstructered text. Once useful information are extracted, it is easy to enclose them within tags and make them machine-readable this way.

Therefore the goal of this work is to automatically extract ingredients with their quantity and unit from recipes and apply tags arround them for further processing.

1.2 Practical usefulness

Beside that text mining is interesting from a computer science point of view, beeing able to extract automatically ingredients from recipes and build a huge machine-readable data pool, can help in many scenarios.

The most obvious one is for cooking-services. It seems, that there is a huge cooking community with sheer endless of recipes for every region like e.g. allrecipes[1]

[1] http://allrecipes.com/

for English, CHEFKOCH.DE[2] for German and cookpad[3] for Japanse speaking regions. They all have the advantage over traditional cookbooks, that user can share their expierence with recipes as well as their modifications. Furthermore they are the basis for recommodation networks for recipes. There are many studies, which take these sites as data pool and recommend recipes for example based on favorite ingredients or nutritionally value calculated out of the ingredients within a recipe (e.g. Teng et al. 2012 or Ueda et al. 2011).

Having a huge machine-readable base of recipes and its ingredients can also provide insights in sociological points of interest. For example in XXX is a comparison between west and asia recipe cuisines.

There are many more interesting questions, which could be analysed with the help of such a huge data pool like a historic analysis of the development and changes of cooking. Occurences of non-local ingredients or meals are evidence for intercultural exchange and globalisation. More expensive ingredients could be an indication for prosperity, while very simple kitchen for poverty or even wartimes.

Therefore I hope, that the data pool of recipes and its ingriedients made in this work, will be helpful for further studies.

1.3 Structure of this thesis

Unknown yet :P

[2] http://www.chefkoch.de/

[3] http://cookpad.com/

Chapter 2

Theory

2.1 Machine-readable data and Ontologies

2.2 Text mining

- was first mentioned in (Feldman and Dagan 1995) - overview (Hotho et al. 2005) -recipe parsing problem and structured prediction problem

2.3 Exact formulation of the problem

2.4 Similar work and distinction to this thesis

Teng et al. 2012 or Ueda et al. 2011 (semi struktur -¿ RE sind gut genug)

2.5 Used algorithms

Chapter 3

Our recipe parser

3.1 Data preparation

3.2 Algorithm

Chapter 4

Evaluation

4.1 Our Website

4.2 Some more or less awesome metrics and formulas

Chapter 5

Summary

Appendix A

Statutory Declaration

I declare that I have developed and written the enclosed Master Thesis completely by myself, and have not used sources or means without declaration in the text. Any thoughts from others or literal quotations are clearly marked. The Master Thesis was not used in the same or in a similar version to achieve an academic grading or is being published elsewhere.

Location, Date Signature

Appendix B

Something else

hi

Appendix C

Something else else

hello

Bibliography

Berners-Lee, T., J. Hendler, and O. Lassila (2000). "Semantic web". In: *Scientific American, 1(1):68–88.*

Erdmann, M. et al. (2000). "From Manual to Semi-automatic Semantic Annotation: About Ontology-based Text Annotation Tools". In: *Scientific American, 1(1):68–88.*

Feldman, Ronen and Ido Dagan (1995). "Knowledge Discovery in Textual Databases (KDT)". In: *Proceedings of the First International Conference on Knowledge Discovery and Data Mining.* KDD'95. Montréal, Québec, Canada: AAAI Press, pp. 112–117. URL: `http://dl.acm.org/citation.cfm?id=3001335.3001354`.

Hotho, Andreas, Andreas Nürnberger, and Gerhard Paaß (2005). "A brief survey of text mining". In: *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology.*

Schema.org (2016). *Schema.org/Recipe.* URL: `https://schema.org/Recipe`.

Teng, Chun-Yuen, Yu-Ru Lin, and Lada A. Adamic (2012). "Recipe Recommendation Using Ingredient Networks". In: *Proceedings of the 4th Annual ACM Web Science Conference.* WebSci '12. New York, NY, USA: ACM, pp. 298–307. ISBN: 978-1-4503-1228-8. DOI: `10.1145/2380718.2380757`. URL: `http://doi.acm.org/10.1145/2380718.2380757`.

Ueda, Mayumi, Mari Takahata, and Shinsuke Nakajima (2011). "User's Food Preference Extraction for Personalized Cooking Recipe Recommendation". In: *Proceedings of the Second International Conference on Semantic Personalized Information Management: Retrieval and Recommendation - Volume 781.* SPIM'11. Aachen, Germany, Germany: CEUR-WS.org, pp. 98–105. URL: `http://dl.acm.org/citation.cfm?id=2887675.2887686`.