Ontology Matching
- using -
Outlier Detection

Master Thesis

presented by
Alexander Müller
Matriculation Number 1376818

submitted to the
Chair of Information Systems V
Prof. .Dr. Heiko Paulheim
University Mannheim

Mai 2015

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overcoming the Disparate Data Space

1.2 Motivation

1.3 Contributions

Chapter 2

The Ontology Matching Problem

2.1 Definition

Formal Definition of ontology matching

2.2 State-of-the Art

What is the current state in research, e.g. current advances at OAEI

2.3 Challenges

Basicalle the Euzenant Paper

Chapter 3

Ontology Matching Approaches (Related Work)

This chapter presents the results of a conducted literature survey in the area of ontology matching.

3.1 Classification of Approaches

Tailored classification of approaches

3.2 Base Matcher

3.2.1 Label Based

TODO

3.2.2 Instance Based

TODO

3.2.3 Structure Based

TODO

3.3 Hybrid Matching Approaches

3.4 Analysis of Hybrid Matching Approaches

Show weaknesses of current approaches, supervised, often weighted average based, so not flexible (transferable to other domains)

Chapter 4

Hybrid Ontology Matching using Outlier Analysis

4.1 Definition Outlier Analysis

4.2 Motivation for using Outlier Analysis for Ontology Matching

Flexibility towards changing data domains No need to train weights upfront

4.3 Ontology Matching as an Outlier Detection Problem

4.3.1 Creating the Feature Vector

4.3.2 Significance of Outliers for Ontology Matching

4.3.3 Transforming the Outlier Analysis Result to a Matching

Chapter 5

A Matching Pipeline using Outlier Detection

5.1 Overview

Presents the implemented Pipeline

5.2 Base Matcher used

What base matcher survived the selection process

5.3 Methods used to combine Matcher

5.4 Feature Selection

5.5 Outlier Analysis

Chapter 6

Evaluation

6.1 Datasets

6.2 Experimental Setup

6.3 Used Baselines

6.4 Results

Chapter 7

Discussion

7.1 Flexibility towards changing Data Domains

7.2 Runtime Considerations

7.3 Comparison with current OAEI Participants

Chapter 8

Conclusion

Bibliography

[1] http://commoncrawl.org/.

[2] http://webdatacommons.org/.

[3] http://webdatacommons.org/webtables/index.html.

[4] Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 136–145. Springer Berlin Heidelberg, 2002.

[5] C. Batini, M. Lenzerini, and S. B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Comput. Surv.*, 18(4):323–364, December 1986.

[6] Philip A. Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching with cupid. Technical Report MSR-TR-2001-58, Microsoft Research, August 2001.

[7] Alexander Bilke and Felix Naumann. Schema matching using duplicates. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 69–80. IEEE, 2005.

[8] Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.

[9] Michael J. Cafarella, Alon Halevy, and Nodira Khoussainova. Data integration for the relational web. *Proc. VLDB Endow.*, 2(1):1090–1101, August 2009.

[10] Michael J. Cafarella, Alon Halevy, Daisy Zhe Wang, Eugene Wu, and Yang Zhang. Webtables: Exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1):538–549, August 2008.

[11] William Cohen, Pradeep Ravikumar, and Stephen Fienberg. A comparison of string metrics for matching names and records. In *KDD Workshop on Data Cleaning and Object Consolidation*, volume 3, pages 73–78, 2003.

[12] William W Cohen. Integration of heterogeneous databases without common domains using queries based on textual similarity. In *ACM SIGMOD Record*, volume 27, pages 201–212. ACM, 1998.

[13] Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy, Hongrae Lee, Fei Wu, Reynold Xin, and Cong Yu. Finding related tables. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 817–828, New York, NY, USA, 2012. ACM.

[14] Hong-Hai Do and Erhard Rahm. Coma: a system for flexible combination of schema matching approaches. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 610–621. VLDB Endowment, 2002.

[15] Songyun Duan, Achille Fokoue, Oktie Hassanzadeh, Anastasios Kementsietsidis, Kavitha Srinivas, and Michael J Ward. Instance-based matching of large ontologies using locality-sensitive hashing. In *The Semantic Web–ISWC 2012*, pages 49–64. Springer, 2012.

[16] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition, 2013.

[17] C. Fellbaum. *WordNet: An Electronic Lexical Database*. Language, speech, and communication. MIT Press, 1998.

[18] W. N. Francis and H. Kucera. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[19] Bernardo Cuenca Grau, Zlatan Dragisic, Kai Eckert, Jérôme Euzenat, Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jiménez-Ruiz, Andreas Oskar Kempf, Patrick Lambrix, et al. Results of the ontology alignment evaluation initiative 2013. In *Proc. 8th ISWC workshop on ontology matching (OM)*, pages 61–100, 2013.

[20] Toni Gruetze, Christoph Böhm, and Felix Naumann. Holistic and scalable ontology alignment for linked open data. In *LDOW*, 2012.

[21] Prateek Jain, Pascal Hitzler, Amit P Sheth, Kunal Verma, and Peter Z Yeh. Ontology alignment for linked open data. In *The Semantic Web–ISWC 2010*, pages 402–417. Springer, 2010.

[22] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[23] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, SIGDOC '86, pages 24–26, New York, NY, USA, 1986. ACM.

[24] VI Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, 1966.

[25] Wen-Syan Li and Chris Clifton. Semantic integration in heterogeneous databases using neural networks. In *VLDB*, volume 94, pages 12–15, 1994.

[26] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[27] Sabine Maßmann, Salvatore Raunich, David Aumüller, Patrick Arnold, and Erhard Rahm. Evolution of the coma match system. In *OM*, 2011.

[28] Erhard Rahm. Towards large-scale schema and ontology matching. In Zohra Bellahsene, Angela Bonifati, and Erhard Rahm, editors, *Schema Matching and Mapping*, Data-Centric Systems and Applications, pages 3–27. Springer Berlin Heidelberg, 2011.

[29] Erhard Rahm and Philip A Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

[30] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[31] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975.

[32] Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. Paris: Probabilistic alignment of relations, instances, and schema. *Proc. VLDB Endow.*, 5(3):157–168, November 2011.

[33] Mikalai Yatskevich and Fausto Giunchiglia. Element level semantic matching using wordnet. In *Meaning Coordination and Negotiation Workshop, ISWC*, 2004.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelorarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Mannheim, den 31.5.2015 Unterschrift