The Extremes of Good and Evil

Master Thesis

presented by
Earl Hickey
Matriculation Number 9083894

submitted to the
Data and Web Science Group
Prof. Dr. Right Name Here
University of Mannheim

August 2014

Contents

List of Algorithms

List of Figures

List of Tables

Chapter 1

Phase I - Data Translation

If you cite something, do it in the following way.

- Conference Proceedings: This problem is typically addressed by approaches for selecting the optimal matcher based on the nature of the matching task and the known characteristics of the different matching systems. Such an approach is described in [3].
- Journal Article: S-Match, described in [2], employs sound and complete reasoning procedures. Nevertheless, the underlying semantic is restricted to propositional logic due to the fact that ontologies are interpreted as tree-like structures.
- Book: According to Euzenat and Shvaiko [1], we define a correspondence as follows.

These are some randomly chosen examples from other works. Take a look at the end of this thesis so see how the bibliography is included.

1.1 Use Case & Data Profiling

1.2 Consolidated Schema & Transformations

Chapter 2

Phase II - Identity Resolution

2.1 Gold Standard

In order to create the gold standard we ran initial identity resolutions with cheap matching rules. With a threshold of 0.3 we use Jaccard 3 Grams, Levensthein Similarity and a combined matching rule (check IDs)!! File Gold Standard integration. The results were then combined into one file which for each individual correspondence outline the similarity calculated by every similarity measure as well as the company names. Correspondences were labeld as certain matches if at least one of the similarity scores had a high matching threshold ¿ 0.95 and an actual match was present. Correspondences were labeled as fuzzy if the similarity measures did not agree on a rating which was indicated by the average similarity. Correspondecnes with low average similarity were labeled as obvious non matches after verifying that they indeed do not match. Afterwards a random sample out of every category was drawn so that the distribution 20/30/50 distribution for the gold standard was being met. The data was then split into train and test set with an sklearn python script, with a test size of ... and stratified on the gold standard category.

–Note on balance of gold standard

2.2 Matching Rules

2.3 Blockers

2.4 Analysis of Errors

Chapter 3

Phase III - Data Fusion

	Baselines			Decision Tree			
Ontology	M(edian)	G(ood)	E(vil)	results	Δ-M	Δ-G	Δ-E
#301	0.825	0.877	0.877	0.855	+0.030	-0.022	-0.022
#302	0.709	0.753	0.753	0.753	+0.044	+0.000	+0.000
#303	0.804	0.860	0.891	0.816	+0.012	-0.044	-0.075
#304	0.940	0.961	0.961	0.967	+0.027	+0.006	+0.006
Average	**0.820**	**0.863**	**0.871**	**0.848**	**+0.028**	**-0.015**	**-0.023**

Table 3.1: Comparison between the Good and the Evil

Bibliography

[1] Jerome Euzenat and Pavel Shvaiko. *Ontology Matching*. Springer, 2007.

[2] Fausto Giunchiglia, Mikalai Yatskevich, and Pavel Shvaiko. Semantic matching: Algorithms and implementation. *Journal on Data Semantics*, 2007.

[3] Malgorzata Mochol and Anja Jentzsch. Towards a rule-based matcher selection. In *Proceedings of the 16th international conference on Knowledge Engineering: Practice and Patterns*, Acitrezza, Italy, 2008.

Appendix A

Program Code / Resources

The source code, a documentation, some usage examples, and additional test results are available at ...

They as well as a PDF version of this thesis is also contained on the CD

Appendix B

Phase II - Identity Resolution

-ROM attached to this thesis.

Appendix C

Further Experimental Results

In the following further experimental results are ...

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Master-/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 Er- klärung rechtliche Folgen haben wird.

Mannheim, den 31.08.2014 Unterschrift