A Replication Study: Understanding What Drives the Performance in WikiMatch

Lu Zhou and Michelle Cheatham

DaSe Lab, Wright State University, Dayton OH 45435, USA,
{zhou.34, michelle.cheatham}@wright.edu

Abstract. We replicate and demonstrate that the performance of the WikiMatch automated ontology alignment system may be driven not by the particular information from Wikipedia directly used by the system, but rather by string similarity and Wikipedia's manually curated synonym sets, as encoded in the site's query resolution and page redirection system. In order to gain a detailed understanding of how Wikipedia contributes to WikiMatch, we replicate results reported for WikiMatch and analyze the results to evaluate our hypothesis.

1 Introduction

This paper reviews an ontology alignment system called WikiMatch. We attempt to replicate the results of the system in order to understand how Wikipedia contributes to its performance. Additionally, we conduct experiments to analyze where the performance comes from. We find that using Wikipedia can in fact find more non-syntactic pairs then using only string similarity. However, the results showed that the performance on both the conference and anatomy datasets were driven primarily by the syntactic similarity of entity labels and secondarily by the Wikipedia page redirection system.

2 Replication and Analysis

The idea behind WikiMatch is to use Wikipedia's general search functionality (through the MediaWiki API[1]) to retrieve a list of related article titles for each of the entities in the two ontologies to be aligned. After retrieving the list of titles, the similarity of each pair of entities is computed by the Jaccard index[2] on these titles. If the similarity exceeds a threshold, WikiMatch considers the entities equivalent. We began our WikiMatch replication effort by downloading the source code from the link specified in [1]. We were able to compile and run the code with minimal effort, and our results were very similar to those in the [1]. Then we used two different datasets: the conference track and anatomy track from the OAEI[3] to explore the factors driving the performance of the system.

[1] https://www.mediawiki.org/wiki/API:Search

[2] https://en.wikipedia.org/wiki/Jaccard_index

[3] http://oaei.ontologymatching.org/

Dataset	Features	Precision	Recall	F-measure	TP	FP	FN
Conference	Levenshtein String Similarity(Baseline)	0.74	0.49	0.58	150	52	155
	Directed + Redirected Queries	0.74	0.49	0.58	150	52	155
	WikiMatch(Directed + Redirected + Article Titles)	0.70	0.50	0.58	152	64	153
Anatomy	Levenshtein String Similarity(Baseline)	0.99	0.62	0.77	937	11	579
	Directed + Redirected Queries	0.99	0.62	0.77	947	11	569
	WikiMatch(Directed + Redirected + Article Titles)	0.96	0.64	0.77	966	43	550

Table 1: Comparison of different approaches on the OAEI conference (Line 1-3) and Anatomy (Line 4-6) Track (TP = True Positives, FP = False Positives, FN = False Negatives, Directed = Identical Terms with Same Title List, Redirected = Different Terms with Same Title List, Article Titles = Different Terms with Different Title List)

Table 1 shows the performance of WikiMatch compared with two other approaches to ontology alignment on two datasets. The first row of each dataset shows the performance achieved by considering two entities equivalent if their labels have a Levenstein string similarity above a threshold of 0.95. The second row shows the performance achieved by considering two entities to be equivalent if querying Wikipedia for them returns the same article. This is possible even when the entity labels are not identical because every article in Wikipedia has a *primary* term associated with it, as well as zero or more *secondary* terms that redirect to that article. For example, the primary term associated with the article on the United States of America is "United State of America", while secondary terms include "United States of America", "America", "US", and "USA". So, "United States of America" in one ontology would be found equivalent to "USA" in another ontology through this method. The final row shows the performance of the full WikiMatch system. Note that WikiMatch performs a *general search* of Wikipedia, meaning that if no article has the search term as a primary or secondary term, the search will continue over the article contents.

Overall, the percentages of correctness from string matching in the conference and anatomy dataset are 98.7% (150/152) and 98.1% (947/966) respectively. These results show that the performance of WikiMatch is mainly driven not by the article titles from Wikipedia that were used, but rather by equivalent labels string matching and the Wikipedia redirection system.

Acknowledgments This work was supported by the US Geological Survey agreement G16AC00120 "Demonstration of Semantic Web Technologies as Applied to Surface Water Feature Classification."

References

1. Hertling, S., Paulheim, H.: Wikimatch: using wikipedia for ontology matching. In: Proceedings of the 7th International Conference on Ontology Matching-Volume 946. pp. 37–48. CEUR-WS. org (2012)