Zhishi.links Results for OAEI 2011

Xing Niu, Shu Rong, Yunlong Zhang, and Haofen Wang

APEX Data & Knowledge Management Lab
Shanghai Jiao Tong University
{xingniu, rongshu, zhangyunlong, whfcarter}@apex.sjtu.edu.cn

Abstract. This report presents the results of **Zhishi.links**, a distributed instance matching system, for this year's Ontology Alignment Evaluation Initiative (OAEI) campaign. We participate in Data Interlinking track (DI) of IM@OAEI2011. In this report, we briefly describe the architecture and matching strategies of Zhishi.links, followed by an analysis of the results.

1 Presentation of the system

Ontology matching is a positive effort to reduce heterogeneity of semantic data. Both schema-level matching and instance-level matching aim at finding matches between semantically related entities of different ontologies. With the development of Linked Data[1], the needs of finding high quality `<owl:sameAs>` links are increasing. Thus, more and more people are engaged in research of identifying URIs referred to the same real-world object by using instance matching techniques.

Zhishi.links we proposed here is an efficient and flexible instance matching system, which utilizes distributed framework to index and process semantic resources, and uses scalable matching strategies for calculating similarities between two resources.

1.1 State, purpose, general statement

Zhishi.links is used for participating in Data Interlinking track (DI) of IM@OAEI2011. Its architecture is shown in Figure 1.

All the dumps (DBpedia, Freebase and GeoNames) are downloaded beforehand and the interconnection is done locally, even the datasets are very large. Admittedly, using keyword search or lookup services provided by these three chosen online database can help to obtain high quality match candidates, but this strategy relies too much on the retrieval performance. Moreover, a large quantity of datasets in Linked Data do not provide such kind of services.

In order to dramatically reduce the time complexity of matching procedures on large datasets, resources are indexed before a pipe of similarity calculations are performed. This principle is also adopted by Silk[6], a well-known links discovery framework. Resources are usually indexed by terms of their names and aliases. In other words, resource pairs sharing more than one term are treated as match candidates and wait for further checking.

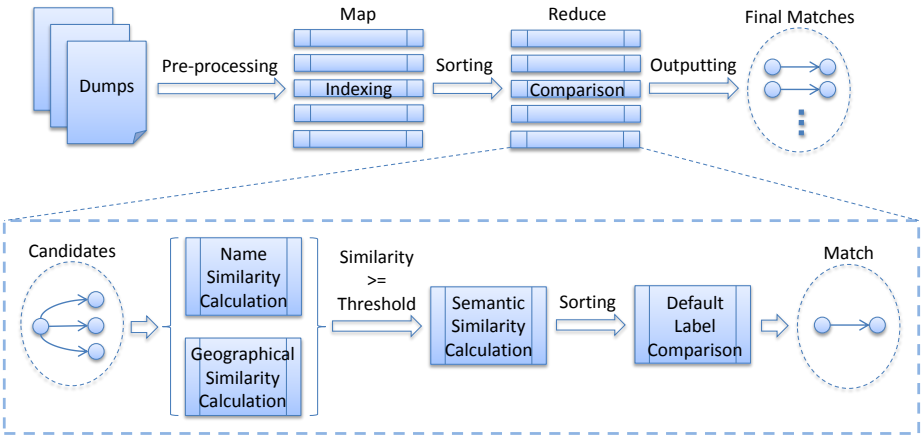

Fig. 1. Architecture of Zhishi.links

The comparison between two resources begins from the string similarity calculation. The more terms two resources' names share, the more similar they are. If a resource has more than one name (i.e. it has aliases), all names take part in similarity calculations and the highest similarity value is chosen. String similarities are used to filter out the least likely match candidates by setting a proper threshold.

Afterwards, semantic similarities are calculated. Generally speaking, similarity scores computed in the previous step increases if two resources have some sematic resemblances (e.g. the same property-value pairs), otherwise penalties are paid. Specifically, functional properties (`owl:FunctionalProperty`) and inverse-functional properties (`owl:InverseFunctionalProperty`) have higher weights than ordinary properties.

Finally, match candidates are sorted by their similarity scores. In some cases, two or more match candidates may have the same highest similarity scores. Zhishi.links compares their default labels again and chooses the closest matching pair.

1.2 Specific techniques used

Since these three datasets are huge, even though we adopt index-based pre-matching, we still suffer from high time and space complexities. We utilize distributed MapReduce[3] framework to accomplish this work. The map function produces a list of key-value pairs, where the keys are the index terms and the values are complete semantic descriptions of resources. After sorting, resources with the same index term (match candidates) gather together and further comparisons are made by reduce function in one computing node.

Properties unique to an object, such as inverse-functional properties, can be used to determine its identity[4]. Values of this kind of properties can also be used to filter or generate match candidates. Here we cite an example of using geographical coordinate to filter candidates before semantic similarity calculation. Most of the time, the geographical coordinate is unique to a location, but unfortunately, the data type of co-

ordinates is a pair of floating point numbers and they can not be used as index terms. So as shown in Figure 1, we can calculate geographical similarity by computing distance between two coordinate. Intuitively, if two locations are wide apart, the corresponding match candidate should be filtered out.

1.3 Adaptations made for the evaluation

In Data Interlinking track, participates are asked to retrieve New York Times interlinks with DBpedia, Freebase and GeoMames. Structured data provided by New York Times is relatively scant. Usually, New York Times just provides a resource's name using `skos:prefLabel` property[1]. In some cases, a short definition (`skos:definition`) and a topic page's URL (`nyt:topicPage`) can be obtained. Thus, Virtual Documents are constructed for resources from other three data sources by splicing values of characteristic properties. Similarity between a Virtual Document and a topic page (abstract, articles' titles and snippets are included) is calculated in semantic similarity calculation phase.

Nearly all default names and aliases in these four data sources are well-designed. Many of them are appended disambiguation information (e.g. "Michael Mann (director)") or supplements (e.g. "University of California, Los Angeles"). Such phrases are isolated because 1. they can be treated as values of characteristic properties and used to calculate semantic similarities, and 2. they may bring about noise when the complete labels are used for string similarity calculation.

Beside these appended phrases, several special words in names are extracted for producing unified values of characteristic properties. For resources which are instances of "People", "Jr." and "Sr." are detected due to the reason that these words have good discriminability. For instances of "Locations" and "Organizations", more keywords are concerned and the full lists of these keywords are shown in Appendix section.

The "name" and "alias" we mentioned above refer to different properties in different data sources. Table 1 shows the exact properties that are used as "name" and "alias".

Table 1. Properties Used as "Name" and "Alias"

Data Source	Name	Alias
New York Times	`skos:prefLabel`	—
DBpedia	`rdfs:label`	`dbpedia-owl:wikiPageRedirects`
Freebase	`rdfs:label`	`fb:common.topic.alias`
GeoNames	`gn:name`	`gn:alternateName`

1.4 Link to the system and parameters file

The homepage of Zhishi.links is `http://apex.sjtu.edu.cn/apex_wiki/Zhishi.links`. More information about our instance matching system can be found here.

[1] Results of using `nyt:search_api_query` are unreachable for us.

1.5 Link to the set of provided alignments (in align format)

The results of Zhishi.links for IM@OAEI2011 can be found at `http://apex.sjtu.edu.cn/apex_wiki/Zhishi.links`.

2 Results

In this section, we will present the result of Zhishi.links for DI track in detail and give related analysis. Tests were carried out on a Hadoop computer cluster. Each node has a quad-core Intel Core 2 processor (4M Cache, 2.66 GHz), 8GB memory. The number of reduce tasks was set to 50.

2.1 DI-nyt-geonames

The version of GeoNames dump we used is May 4th.

The final results are shown in Table 2, where precision, recall and f-measure are provided. After filtering, 128,795 match candidates were sent to semantic similarity calculation component and approximately 98.9% (H_Recall) expected matches were include.

One reason for the notable decrease of recall is that URI aliases exist in this dataset: different URIs are used to identify the same location. For example, `gn:1863967`[2] and `gn:1863961`[3] refer to the same place in Japan: their names are both "Fukuoka" and their geographical coordinates are very close. We can hardly resolve this problem off-line except some official rules are provided.

Table 2. Performance of Zhishi.links (NYT-GeoNames)

Type	Precision	Recall	F-measure	Candidates	Expected	H_Recall
Locations	0.938	0.883	0.910	128,795	1,789	0.989

2.2 DI-nyt-dbpedia

The version of DBpedia dump we used is 3.6.

Zhishi.links performs best on DBpedia, as shown in Table 3. Wikipedia's strict and advanced naming conventions[4] and abundant aliases guarantee the quality and quantity of resources' names separately, which help a lot in similarity calculation phase.

However, precisions here are not satisfactory. We have investigated the causes and found that many incorrect matches occurred when ambiguities were existing. As explained in Section 1.3, New York Times does not provide sufficient structured descriptive data, hence more sophisticated matching methods should be applied.

[2] http://sws.geonames.org/1863967/

[3] http://sws.geonames.org/1863961/

[4] `http://en.wikipedia.org/wiki/Wikipedia:Naming_conventions`

Table 3. Performance of Zhishi.links (NYT-DBpedia)

Type	Precision	Recall	F-measure	Candidates	Expected	H_Recall
People	0.971	0.970	0.970	16,787	4,977	0.992
Organizations	0.896	0.932	0.913	10,679	1,965	0.957
Locations	0.910	0.914	0.912	47,490	1,920	0.983

2.3 DI-nyt-freebase

The version of Freebase dump we used is July 7th.

URI aliases also exist in Freebase. RDF dump built by RDFizer[5] does not contain URIs for resource. They should be generated by using values of `fb:type.object.key`. Unfortunately, this procedure produces more than one URI for a single resource. For example, `http://rdf.freebase.com/ns/en.amanda_hesser`, `http://rdf.freebase.com/ns/user.jamie.nytdataid.6389285617816563261 3`, `http://rdf.freebase.com/ns/wikipedia.en.Amanda_Hesser`, etc. all refer to a person named "Amanda Hesser".

Unlike GeoNames, only one URI for a resource is chosen in pre-processing phase. That is why the highest recalls (H_Recalls), as shown in Table 4, are also unsatisfactory. The priority list we used is:

1. http://rdf.freebase.com/ns/en.*
2. http://rdf.freebase.com/ns/user.jamie.nytdataid.*
3. http://rdf.freebase.com/ns/authority.us.gov.loc.na.n*
4. http://rdf.freebase.com/ns/authority.iso.*
5. http://rdf.freebase.com/ns/business.cik.*

If the URIs in reference alignments do not follow this priority list, mistakes would be unavoidable.

Table 4. Performance of Zhishi.links (NYT-Freebase)

Type	Precision	Recall	F-measure	Candidates	Expected	H_Recall
People	0.929	0.924	0.926	26,382	4,979	0.964
Organizations	0.887	0.853	0.870	12,664	3,044	0.889
Locations	0.902	0.865	0.883	14,705	1,920	0.932

3 General comments

In this section, we will give some additional comments on Zhishi.links results and provide some suggestions to OAEI organizers.

[5] `http://code.google.com/p/freebase-quad-rdfize/`

3.1 Discussions on the way to improve the proposed system

Several shortcomings of Zhishi.links can be seen and need to be overcome in the future:

- When it comes with the problem of homonyms, instance matching systems should exploit as much information as possible to enhance the discriminability of their matchers. Currently, subject to the fact that most descriptions given by New York Times are written in natural language, the performance of our semantic similarity calculator are constrained. We are considering more tests carrying out on datasets in different styles and designing a more robust system.
- In DI track, only three types of resources are involved. The special words in names, which are extracted as values of characteristic properties, are chosen manually. Some smarter strategies should be applied to accomplish this mission.

3.2 Comments on the OAEI 2011 test cases

We are very interested in testing our matching system on large-scale real-world data. It can help validating the robustness and applicability of the proposed methods. However, crude raw data may have some defects. The URI aliases problem, for example, is what we have met. We hope that these issues are resolved in the future or considered in the evaluation.

3.3 Proposed new measures

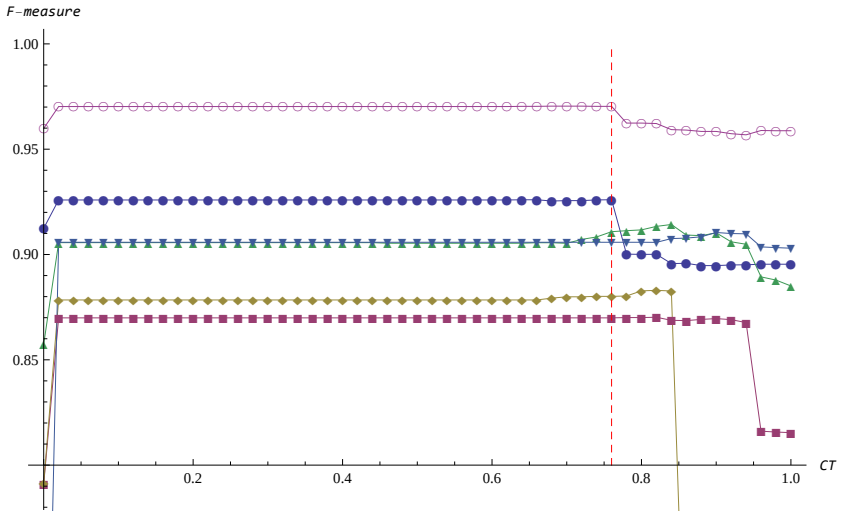

Fig. 2. F-measures on Different Confidence Thresholds

Our goal is to design a general instance matching system. It should not sensitive to what type of resources should be matched and what data source resources come from. So what we need is a matching method with high stability.

As Shown in Figure 2, we choose match candidates above continuously varying *confidence threshold*s (*CT*) and plot the corresponding *F-measure*s on the chart. The six curves in this chart indicate the results for matching tasks carried out on DBpedia and Freebase. We can determined a fixed *CT* value to filter final match candidates. For instance, we can choose $CT = 0.76$ here. Then the performances are not the best, but are relative acceptable (other matching systems are not taken into consideration).

Here we just mentioned the concept of **stability**. The complete descriptions for evaluating the stability of matching systems are elaborate in [5].

4 Conclusion

In this report, we have presented a brief description of Zhishi.links, an instance matching system. We have introduced the architecture of our system and specific techniques we used. Also, the results have been analyzed in detail and several guides for improvements have been proposed. We look forwards to build an instance matching system with better performance and higher stability in the future.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5(3), 1–22 (2009)
2. Bouquet, P., Stoermer, H., Tummarello, G., Halpin, H. (eds.): Proceedings of the WWW2007 Workshop I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web, Banff, Canada, May 8, 2007, CEUR Workshop Proceedings, vol. 249. CEUR-WS.org (2007)
3. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: OSDI. pp. 137–150 (2004)
4. Hogan, A., Harth, A., Decker, S.: Performing object consolidation on the semantic web data graph. In: Bouquet et al. [2]
5. Niu, X., Wang, H., Wu, G., Qi, G., Yu, Y.: Evaluating the stability and credibility of ontology matching methods. In: Antoniou, G., Grobelnik, M., Simperl, E.P.B., Parsia, B., Plexousakis, D., Leenheer, P.D., Pan, J.Z. (eds.) ESWC (1). Lecture Notes in Computer Science, vol. 6643, pp. 275–289. Springer (2011)
6. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and Maintaining Links on the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) International Semantic Web Conference. Lecture Notes in Computer Science, vol. 5823, pp. 650–665. Springer (2009)

Appendix

Table 5. Unified Values for Organizations

Keywords	Unified Values
Co, Company	Co.
Corp, Corporation	Corp.
Inc, Incorporated	Inc.
Ltd, Limited	Ltd.

Table 6. Unified Values for Locations

Keywords	Unified Values
North Carolina	NC
New Hampshire	NH
New Jersey	NJ
New York	NY
Bronx, Brooklyn, Manhattan, Queens	NYC
Rhode Island	RI
Florida	Fla
Georgia	Ga
Louisiana	La
Maine	Me
Mississippi	Miss
Missouri	Mo
Pennsylvania	Pa
Virginia	Va
Vermont	Vt