Data Integration Beyond Alignments Between Two Sources

ABSTRACT

Every day, new data sources are becoming available. In data integration, one often seeks to merge two sources by first computing similarity scores between items, and then selecting a 1-to-1 alignment of maximal weight, e.g. via the Hungarian algorithm for bipartite matching.

We can generalize this to more than two sources and even beyond the strict 1-to-1 alignments of k-partite variants of maximum matching. In general, we may have items from any number of sources, with arbitrary weighted links indicating possible identity, as well as one or more groups of sets of items indicating likely distinctness. Within each group of sets, two items in different sets are assumed distinct with some weight. Bipartite matching can then be reduced to the special case of assuming a node is pairwise distinct from all other nodes on the same side. However, this formalism flexibly allows for capturing many other scenarios, and distinctness needn't be a hard constraint.

Making this identity and distinctness evidence consistent, under the standard assumption of transitivity for identity, is NP-hard as well as APX-hard. Still, one can obtain good solutions (with a logarithmic approximation guarantee) using graph flow techniques [3].

It turns out that this has practical applications in a number of areas. These include combining several different sources into a single domain-specific knowledge base [2], finding inconsistent links in the Web of Linked Data [1], and turning the different language-specific editions of Wikipedia into a large integrated multilingual lexicon [3].

BODY

Reconciling identity vs. distinctness of items from different sources can be NP-hard, but approximations enable flexible data integration.

REFERENCES

[1] G. de Melo. Not quite the same: Identity constraints for the Web of Linked Data. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013)*. AAAI Press, 2013.

[2] G. de Melo. Lexvo.org: Language-related information for the Linguistic Linked Data cloud. *Semantic Web*, 6(4):393–400, August 2015.

[3] G. de Melo and G. Weikum. Untangling the cross-lingual link structure of Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*. Association for Computational Linguistics (ACL), 2010.

Volume 4 of Tiny Transactions on Computer Science