Unsupervised Entity Resolution using Graph Summarization

Data Science Institute
COLUMBIA UNIVERSITY

Karan Sindwani, Lisa Sarah Thomas, Neha Saraf, Raj Biswas
Amir Rahmani, Jihan Wei, James Krach
Tian Zheng

Data Science Capstone Project with Capital One

Introduction

Entity Resolution refers to the task of finding all mentions of same real-world entity within a knowledge base or across multiple knowledge bases. In the modern world, the speed and volume of data has increased exponentially. Thus, making inference across networks and semantic relationships between entities a greater challenge to overcome. Entity resolution can reduce the complexity by proposing canonicalized references to entities and deduplicating and linking entities. The applications of entity resolution are tremendous, particularly for public sector and federal datasets related to health, transportation, finance, law enforcement, and antiterrorism.

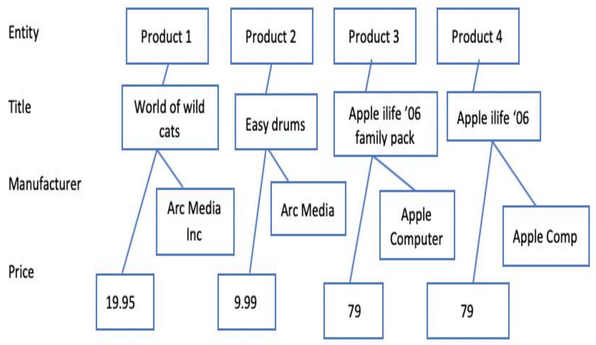

Figure 1. Input graph

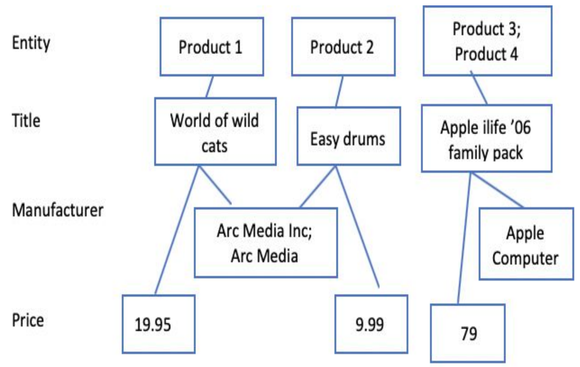

Figure 2. Summary graph

Methodology

We address the problem of performing entity resolution by modeling the data as an RDF graph, shown in Figure 1. In the above graph of products and manufacturers, the goal could be to resolve all mentions of similar products and manufacturers. Using the links between instances of different types and similarity between node contents, we aim to improve the accuracy of the algorithm.

We formulate this problem as a multi-type graph summarization problem, which involves clustering the nodes in each type that refer to the same entity into one super node and creating weighted links among super nodes that summarizes the inter-cluster links in the original graph. The algorithm takes a graph(Figure 1) as input and returns a summary graph as depicted in Figure 2. Experiments show that the proposed approach outperforms several state-of-the-art generic entity resolution approaches, especially in data sets with missing values and one-to-many and many-to-many relations.

Results

For the experiments, we used Amazon-Google products dataset. Using the stated methodology, we computed a summary graph. Figure 3 shows the statistics of the initial graph. The first layer of the graph corresponds to product names forming supernodes of resolved entities. The second layer corresponds to unique words from descriptions. In Figure 4, we can see the algorithm clusters different mentions of product names together.

Data	Types	#records	#nodes	#edges
Google and Amazon Products	2	4589	12397	41165

Figure 3. Statistics of graph

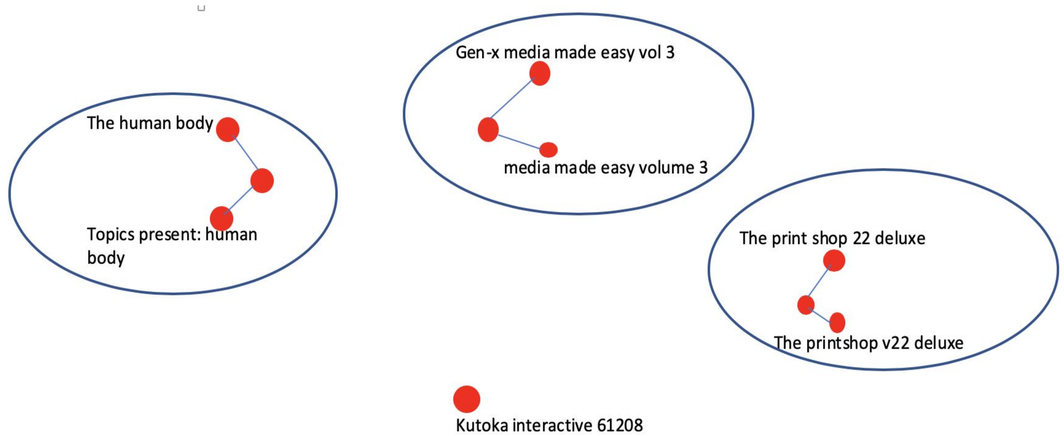

Figure 4. Sample Clustered Products

Conclusion

In this work, we use multi-type graph summarization method that identifies entities in an unsupervised setting. We applied this approach on a dataset of products from Google and Amazon and visually inspected the results. We plan to make a quantitative assessment of this approach and evaluate its performance on other datasets.

Acknowledgments

We are very grateful to Amir Rahmani, Jihan Wei and James Krach from Capital One for many helpful discussions and comments on our approaches and reports. We also gratefully acknowledge Tian Zheng for her support and guidance throughout the process.

References

Unsupervised Entity Resolution on Multi-type Graphs. ISWC 2016 - 15th International Semantic Web Conference. 2016. Linhong Zhu and Majid Ghasemi-Gol and Pedro Szekely and Aram Galstyan and Knoblock, Craig A. http://usc-isi-i2.github.io/papers/zhu16-iswc.pdf