Weighted bipartite matching and its application in data linking

Sina Ahmadi sina.ahmadi@insight-centre.org
Supervisors: John McCrae Mihael Arcan

Insight
Centre for Data Analytics

Introduction

Data linking is used to bring together information from different sources in order to create a new richer data set by identifying and combining information from corresponding entities on each of the different source data sets. One of the main applications of data linking is in **e-lexicography**. In a lexicon, one entry can be linked to one or more entries in another lexicon based on different types of relationship such as synonymy, antonymy or equivalence.
In this study, we are interested in different variants of the bipartite matching problem with a new application in linking cross-lingual lexical data based on similarity scores. We are particularly interested in matching labels and textual description of concepts in Wikipedia where different entries may be juxtaposed to one or more other entries with the same content. Such links are known as **redirects**. Figure 1 shows the redirects to "Galway".

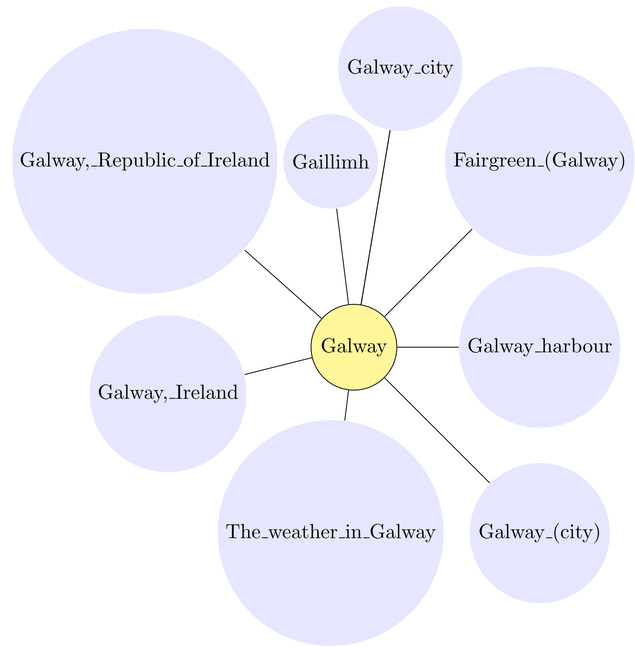

Figure 1: Wikipedia redirected pages to "Galway"

Objective

Our ongoing research focuses on one of the main challenges in e-lexicography which is to find more efficient matching solutions with respect to complex relationships between two sets of entries. We are aiming at analyzing the extent to which graph matching algorithms can improve cross-lingual data linking and, particularly, redirects linking.

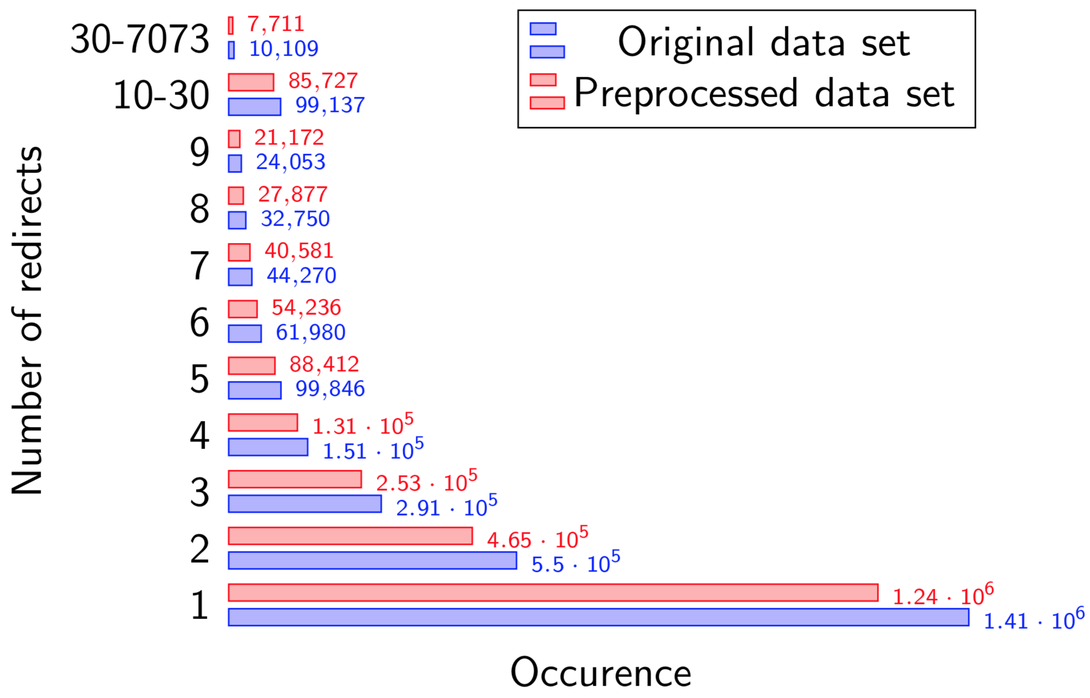

Figure 2: Frequency of redirects in Wikipedia (August 2018)

Figure 2 illustrates proportional amount of redirection links in Wikipedia. Although there is a vast range of frequencies between 1 and 7073, most of the pages that have been redirected to have a frequency of less than 5. Since the target pages are not interconnected, we can model our problem as matching a **bipartite undirected weighted graph** in which the two disjoint sets of vertices are fully connected with edges which are weighted based on **similarity scores**.

Methodology

There are various methods for matching weighted bipartite graphs. One strategy is to select **exhaustively** all edges over a specific threshold. On the other hand, maximum-weight matching suggests an optimal solution as a matching where the sum of weights have a maximal value. This method is known as the **assignment problem**. The Hungarian algorithm is one of the solutions that solves the assignment problem in polynomial time.
Although the maximum weight matching is more efficient than the exhaustive method, it only yields one-to-one links between the vertices. In more recent studies, the **weighted bipartite b-matching (WBM)** algorithm has been shown to be more efficient in matching vertices based on the **capacity** of a vertex. The capacity of a vertex, denoted by b is the number of the vertices that can be matched to that vertex. WBM finds a subgraph which maximizes the sum of the edges having every vertex adjacent to at most b edges.

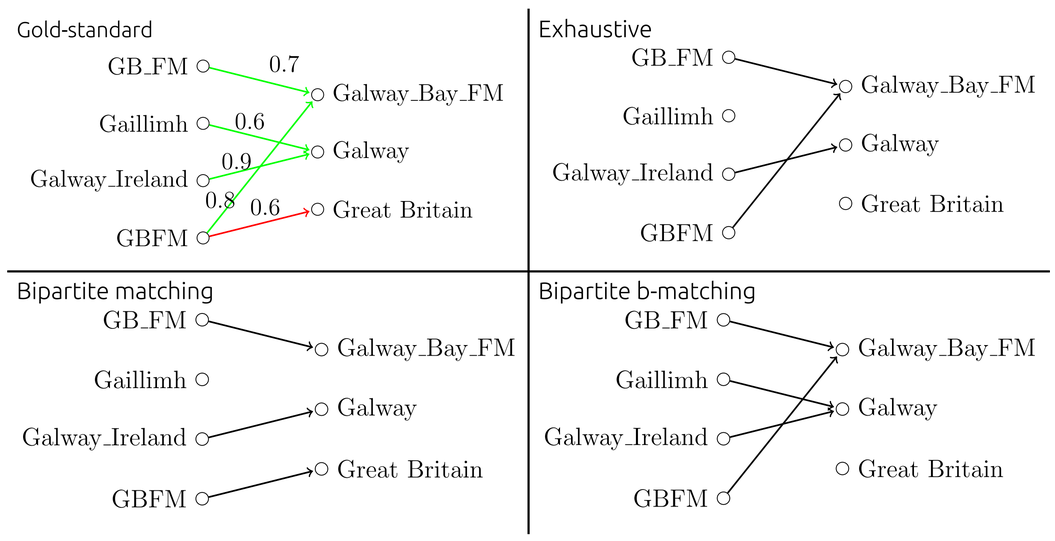

Figure 3: Matching strategies

Experiments

At this stage of our project, our experiments suggest a considerable improvement using the WBM algorithm in terms of precision and recall. Our final results will be reported in an article in the near future.
As a part of the ELEXIS project (`https://elex.is/`), the outcome of our study will be integrated in NAISC, a semi-automated system for link discovery and data linking in lexicography.

A World Leading SFI Research Centre

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.