Implicit translation of Wikipediæ via Random Graph Embeddings

Marchette, David
Naval Surface Warfare Center, Q21
18444 Frontage Rd, Suite 327
Dahlgren, VA 22448,USA
david.marchette@navy.mil

Hohman, Elizabeth
Naval Surface Warfare Center, Q21
18444 Frontage Rd, Suite 327
Dahlgren, VA 22448,USA
elizabeth.hohman@navy.mil

Priebe, Carey
The Johns Hopkins University, Applied Mathematics and Statistics
3400 North Charles Street
Baltimore, MD 21218-2682, USA
cep@jhu.edu

The multilingual Wikipediæ provide a good testbed for developing methods for the analysis of text and the fusion of text and graph information. In this work we focus on the problem of implicit translation: given a Wikipedia in one language L_1 and another in L_2, with some known associations of articles in one to the other, can one determine further associations, without an explicit translation dictionary; for example, determining that the Afrikaans article titled "Sterrekunde" should be matched with the one titled "Astronomie" in Dutch or "Astronomy" in English. To perform the matching, we define a novel random projection which allows us to project the Wikipediæ into the same space, with similarity in this space providing the association.

Suppose we are given two graphs, $G = (V, E_G)$ and $H = (V, E_H)$, on the same vertex set, V. For each vertex $v \in V$, associate a random vector $z_v \in \mathbb{R}^d$. Denote by Z the $|V| \times d$ matrix of these vectors. Denote by d_v the degree of vertex v:

$$d_v = |\{w : vw \in E\}|.$$

Note that if the graphs are directed, this is the out-degree of the vertex. We define the projection of v as an element of the graph G as:

$$\rho(v; G, Z) = \frac{z_v}{d_v} \sum_{vw \in E_G} z_w. \quad (1)$$

Here multiplication of vectors is coordinate-wise. It is worth noting that if we view the projection of an edge vw as $\rho(vw; G, Z) = z_v z_w$, then the vertex is projected to the centroid of its incident edges. More generally, given a weight vector s summing to 1, we can define the projection as:

$$\rho(v; G, Z) = \sum_{vw \in E_G} s_w z_v z_w. \quad (2)$$

Implicit Translation

The Wikipedia in a given language consists of a collection of pages on various topics with links between the pages. This results in a (directed) graph, with each node corresponding to a page. The

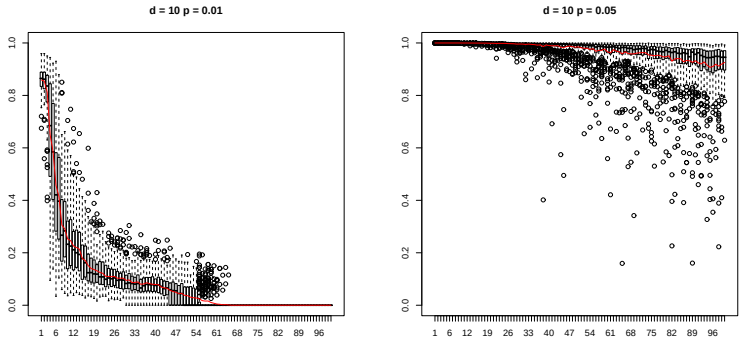

Figure 1: In each plot, the x-axis corresponds to k, the number of edges added and removed from the graph G to construct H. The y-axis is the proportion of non-pair distances greater than the empirical critical value. 100 simulations are run for each value of k. The left plot corresponds to $p = 1/n$ and the left to $p = 5/n$, with $n = 100$ in both. The red curves correspond to the medians.

nodes contain titles and unique identifiers, as well as links to other information such as images and web pages. In addition, for most pages, there are links to "the same" page in a variety of other languages. Thus, we can associate to a given page in one language a page in another. In particular, we will use the fact that English is by far the largest Wikipedia, and so for most pages in any language other than English, there is a corresponding page in English.

Given Wikipedia graphs $G_1, \ldots, G_k$ in languages $L_1, \ldots, L_k$, we associate to each vertex the associated English title, and use these to associate vertices in different languages. We remove vertices which do not have associated English titles. Using the projection method described above, we can project the k graphs into the same space. Now we can compare the distances between paired documents with those between unpaired documents, to determine the extent to which the projection is respecting the pairing.

We explore the performance of this method for pairing Afrikaans and Dutch articles, and Afrikaans and articles written in several Bantu languages. The English associations provided by the Wikipediæ allow an objective performance evaluation. The random embedding method is designed to be applicable to very large graphs, and we demonstrate it on both the small Bantu graphs (about 100 articles each) and the large graphs in Afrikaans (9K articles) and Dutch (322K articles).

Simulations

Consider the following experiment: given an Erdös-Renyí random graph $G(n, p)$, construct a new graph H on the same vertex set by removing k edges and adding k edges. Thus, G and H have the same number of edges, and differ by $2k$ edges. We project both graphs and compute the distances between paired vertices and between unpaired vertices. Fixing $\alpha = 0.05$, we find the critical value c_α such that α percent of the paired vertices have distances greater than c_α and compute the proportion of unpaired distances greater than c_α. Several examples are given in Figure 1.

As can be seen in the figure, for graphs with as few as 5 edges per vertex (on average) there is enough structure retained in the graphs that the power of the hypothesis test (rejecting the null that the documents are paired) is extremely good, even for moderate projection dimension.

A similar test was run using random dot product graphs. (see Marchette and Priebe, 2008). The results, in Figure 2, show that the performance is near perfect for reasonably low projective dimension

Dot Product Graph Simulation

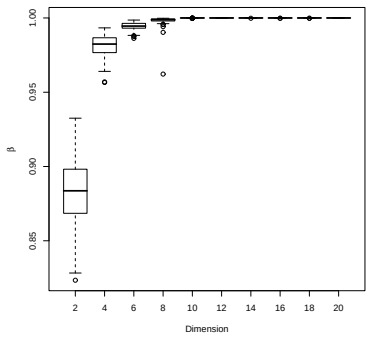

Figure 2: Two random 100-node dot product graphs (with the same vectors) are generated and projected into $\mathbb{R}^d$, as indicated on the x-axis. The y-axis is the proportion of non-pair distances greater than the empirical critical value. 100 simulations were run.

Multilingual Corpus

Table 1: South African language Wikipediæ used in the implicit translation experiment. Only those pages with outgoing links and associated English titles were used in the experiment.

Language	Number of Pages
Afrikaans	9,136
Dutch	322,728
Southern Sotho	28
Swati	112
Tsonga	65
Tswana	60
Venda	41
Xhosa	67
Zulu	56

in this simulation.

Dutch, Afrikaans and Bantu Wikipediæ

We processed the Wikipediæ tabulated in Table 1. There were $325,793$ English titles in total. Each English title corresponds to a vertex label, with all vertices with the same label obtaining the same vector (we chose $d = 10$ for the purposes of illustration). Thus, we can project all the articles of each graph into the same 10-dimensional space.

Figure 3 compares the distance between matched pairs (articles labeled with the same English title in Afrikaans and Dutch) compared to the distance between unmatched articles. Higher dimensional projections have less overlap, however even at the relatively low dimensional projection used here ($d = 10$) there is considerable separation, indicating that for these graphs, implicit translation is possible.

Figure 4 shows an experiment in which each Bantu language Wikipedia is projected into $\mathbb{R}^{100}$. The Afrikaans articles are projected into the same space. For each Bantu article b, the projection is computed as the average of the projections of its neighbors. The distance between b and its paired

Afrikaans vs Dutch

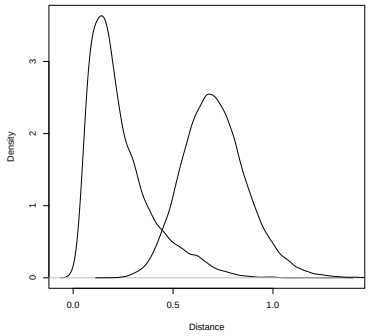

Figure 3: Distances between matched pairs (left curve) and unmatched pairs (right curve) for the Afrikaans and Dutch Wikipediæ.

Bantu Languages

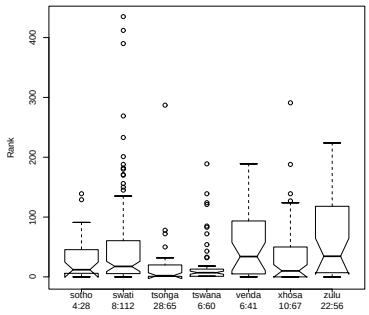

Figure 4: Ranks of the matching pair in the Bantu–Afrikaans test. In each case the Bantu language and Afrikaans Wikipediæ were projected into $\mathbb{R}^{100}$, and the rank of the Afrikaans article matching the Bantu article is computed. The numbers below the language names correspond to the number of isolated vertices and the total number of vertices in the language. Isolated vertices are not matched.

document in Afrikaans is computed, and the rank of this distance among the neighbors of the pairs in Afrikaans is reported. This shows that the implicit translation works, provided that one restricts ones search to "neighbors of neighbors". However, the method cannot apply to isolated vertices. For these, one would have to make use of the text in the documents. For example, one could define a graph based on the similarities of the texts. Further research on the fusion of text and graphs is ongoing.

This work is preliminary. One of the most important issues is out-of-sample embedding. The method chosen here, to embed as the average of the neighbor's projections, works well in this example, but is not well suited to situations in which many vertices need to be embedded at once. There is a problem for which out-of-sample embedding is not necessary: a time series of graphs. In this case, the vertex set is fixed, and the embedded points can be tracked through time. This is an area of current research.

REFERENCES (RÉFERENCES)

D. J. Marchette and C. E. Priebe, "Predicting unobserved links in incompletely observed networks", Computational Statistics and Data Analysis, 52, 1373–1386, 2008.