WIKIPEDIA LINK ANALYSIS

MICHAEL LOTKOWSKI
S1207068@SMS.ED.AC.UK

1. Introduction

Wikipedia is a free-access, free-content Internet encyclopedia, supported and hosted by the non-profit Wikimedia Foundation. Those who can access the site can edit most of its articles. Wikipedia is ranked among the ten most popular websites and constitutes the Internet's largest and most popular general reference work [1]. Like most websites it's dominated by hyperlinks which link pages together. The large size of the English Wikipedia (5,021,159 articles)[2] makes it impractical for analysis in the short time and with limited computational resources. The Scots language Wikipedia has 35,300 articles[2], which makes it a better choice. Using another Wikipedia is a better choice than sampling from the English Wikipedia as the entire structure of the graph is preserved and it is a sufficient size to allow us to believe that the larger Wikipedias behave in a similar manner.

2. Missing Links

The Scots Wikipedia has many articles with very few links. This is mainly due to short article length. However this makes it difficult to navigate on the site. In many cases the article do exist for Scots Wikipedia, however they are just missing a link. The Strongly Connected Component has 25338 out of 44670 nodes, accounting for 56.72% of all the nodes. This will also be true for other small Wikipedias. Wikipedia provides an option to view the same article in many languages. This is a useful option if you know more than one language, however most people would prefer to read an article in the same language. This report presents a method for automatic page linking using a larger Wikipedias. It also explores further improvements to this method.

3. Methodology

For the automatic page linking we require two datasets, a small Wikipedia and a large Wikipedia which we will use to supplement the links for the small Wikipedia. Wikimedia provides database dumps for every Wikipedia [3]. First we need to extract all links from the dumps using Grpahpedia [4] . This will generate a neo4j database, containing all pages as nodes and links as edges. I have also altered Graphpedia to output a json file to make it smaller and easier to manipulate in Python [5]. After the preprocessing is done then we will run analysis on the base graph of the small Wikipedia. Afterwards we will enrich the base graph of the small Wikipedia using the graph structure of the large Wikipedia, and run the same analysis. Lastly we will combine the two graphs together and run the same analysis. The enriched graph will have to be normalised as it's not possible to get new links for all pages.

To evaluate this method I will compare the size of the Strongly Connected Component, average number of links and the number of pages which have no outgoing links. For this method to be successful it has to have a significant increase in interconnectivity and significantly

decrease dead-end pages. The viability of the method will also depend on it's performance.

4. Solution

Wikimedia provides an SQL dump of interlanguage translations. Combined with the page titles dump we can join the two to create a mapping between Scots and English titles. For each Scots title we look up the equivalent English title and search the English Wikipedia for first degree neighbours. For each neighbour node look up the Scots equivalent, if it exists create an edge between the initial title and the neighbour node. Below is a pseudo code outlining this algorithm.

```
Data: SW: small Wikipedia Dump
Data: LW: large Wikipedia Dump
Data: Trans: Mapping between page
      titles
Data: G: New graph
Result: A small Wikipedia graph with
        links mapped from the large
        Wikipedia
for page in SW do
    large_page = Trans.sWToLW(page)
    links = LW.linksForPage(large_page)
    for link in links do
        if link in Trans then
            G[page].addLink(Trans.lWToSW(link))
        end
    end
end
return G
```

Algorithm 1: Automatically generate new links

This algorithm has worst case complexity of $O(nl)$ where n is the number of pages in the small Wikipedia, and l is the number of links in the large Wikipedia. However on average it performs in $\Omega(n)$, as the number of links per page is relatively small, on average about 25 per page in the English Wikipedia [6].

5. Evaluation

Enriched scc is 27366 out of 31516 86.83%
combined scc is 36309 out of 44670 81.28%
avg 34.84

6. Conclusion

7. Further Improvements

References

[1] Wikipedia - wikipedia. https://en.wikipedia.org/wiki/Wikipedia. Accessed: 2015-11-27.

[2] Wikipedia - list of wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias. Accessed: 2015-11-27.

[3] Wikimedia - dumps. https://dumps.wikimedia.org/. Accessed: 2015-11-23.

[4] Github - graphipedia. https://github.com/mirkonasato/graphipedia. Accessed: 2015-11-23.

[5] Github - graphipedia. https://github.com/DownMoney/graphipedia. Accessed: 2015-11-23.

[6] Stephen Dolan - six degrees of wikipedia. http://mu.netsoc.ie/wiki/. Accessed: 2015-11-23.