Slovene Wikipedia Network Analysis

Šimen Ravnik[a,1], **Andrej Kronovšek**[a,2], **and Tim Vučina**[a,3]

[a]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia

The manuscript was compiled on May 28, 2022

Wikipedia is the largest free encyclopedia in the world, created by hundreds of thousands of individuals from all over the world. Wikipedia was originally created in English language, but was then quickly translated into other languages. In the year 2002, Slovene Wikipedia was created and since then around 176 000 articles were published. The Wikipedia's core idea is based on hyperlinks. Hyperlinks are the essence of the internet since they provide quick and efficient information transmission. Our goal was to construct a network from articles published on Slovene Wikipedia which will serve as a ground truth for our project. More precisely, we constructed a network where links represent *wikilinks* (internal links) between articles. Together with this information we also obtained data about the number of views of all pages and the categories the article are linked to. In the first part we performed exploratory data analysis to get a large overview of the network that we are dealing with and our interesting findings are gathered in this report. After the thorough analysis we began tackling two main problem of Wikipedia, which are link prediction and article clustering.

Problem definition, motivation, background, contributions, etc.

The core motivation for our project is to help the community of volunteers that are building Wikipedia regularly in a way, that we improve the information flow and create Wikipedia that is as homogeneous as possible. When thousands of volunteers are writing their articles on Wikipedia, they are dealing with the problem where they might **forgot to reference** an important wiki page that should be included in their article. We want to have an encyclopedia that is as precise and as informative as possible and we achieve that with referencing the right articles. Our goal was to tackle this problem with **link prediction** methods, where we would recommend which *wikilinks* an author might include in the article. The second problem the volunteers on the Wikipedia are dealing with is the categories. If we focus on the categories of the pages in Wikipedia, we can quickly realise that they are extremely messy. The reason for that is that anyone can create a subcategory, meaning that throughout the years lots of categories were created. To be precise, **66.288** categories exists in Slovene Wikipedia, and many of them does not make any sense. Therefore, our goal was to find 5-10 reasonable categories in which the articles are clustered.

But before diving into the prediction models, our first task was to retrieve the data and understand the network that we are dealing with. To retrieve the network we used Wikimedia API, where we could get all the pages (titles and IDs) and the pages they are linking to. With that we constructed the network with **176.413 nodes** and **8.933.180 edges**. Of course the constructed network is a directed network since the pages doesn't necessarily link in both directions. Additionally we also gathered the number of views of the pages for every day in the last two months and the their assigned categories.

Related work

Some network science work has been done on Wikipedia articles, but their applications were for different purposes and they were mainly focusing on the English Wikipedia. Authors of article (1) have tackled collective behaviour analysis on Wikipedia, meaning which pages are most viewed in certain time frame. We also performed similar approach in our exploratory network analysis, because that gave us the core understanding and comparison measure for our pages. Next, some research has been done in measuring article quality using collaboration network (2). The challenge of organizing a set of domain-specific terms into a hierarchical structure was solved using so-called DF-Miner (3). Authors of article (4) were also interesting in the role of diversity in Wikipedia, more specifically, the correlation of teams and the quality of the article. In (5) a temporal approach was taken and the authors analysed the evolution of Wikipedia with 9 different languages, (6) combined the Wikipedia links with natural language processing methods and (7) focused only on science articles of Wikipedia.

Methods & Results

A. Exploratory network analysis. Before starting any kind of modeling we first have to deeply understand the network we are dealing with. Therefore we started our research with analysing the structure of our network. We were mainly focus here on how to visualise the network in a way, that we can quickly see interesting information which will help us improve our prediction models.

We first focused on the degree distribution of our network. We plotted the degree distributions for indegree, outdegree and those two combined, which we can see in Figure 1.

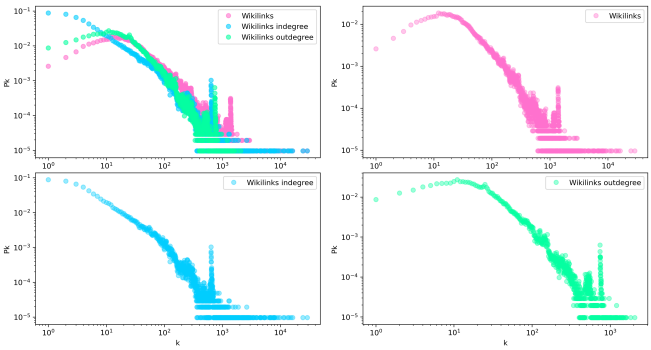

Fig. 1. Degree, indegree and outdegree distribution for Wikilinks network.

From the distributions we can quickly see that the inde-

All authors contributed equally to this work.

[3]To whom correspondence should be addressed. E-mail: tv3843@student.uni-lj.si.

gree of our network follows the power-law and is scale-free, while this property is not completely present in the other two distributions. This was also expected since the importance of the articles on Wikipedia varies, meaning that there exists articles that many other articles cites. Therefore, the hubs are created and consequently the network becomes scale-free. But we can see an interesting spike in all distributions around $k = 1000$. This spike cannot be a result of some randomness. At first we thought there might be some limit when creating links to other pages, but after the examination we realised that pages with approximately 1000 links are mainly year pages (say 1981 or 1948) – for outdegree, and Named Entities (Places, People, etc.) – for indegree, which are all around the same length. That is why a spike is clearly visible. We can see main characteristics of Wikilinks network in Table 1.

Table 1. Main characteristics of Wikilinks network.

	n	m	$\langle k \rangle$	k^{in}_{max}	k^{out}_{max}	$\langle d \rangle$
Wikilinks	176.413	8.933.180	101.3	41.095	4.687	4.31
Random	176.413	8.933.180	101.3	85	86	3.46

Next, we were interesting in how out network actually looks, meaning if there are any communities formulating, which articles are similar to one another, which articles are most important and so on. But since our network is extremely large, we could not use standard approaches when visualizing the network. Our goal was to plot the network in a way, that we keep the relations between the nodes and consequently the structure of the network would be visible. The idea was to transform the nodes into vectors using **node2vec** algorithm which encodes the structure of the network and keep similar articles together in the embeddings space. We implemented the procedure and retrieve a network shown in Figure 2. We determined size and the color of the nodes using **PageRank** algorithm. From the visualization we can quickly see three threads that exist in out network. Everything starts from one point and than starting to spread around.

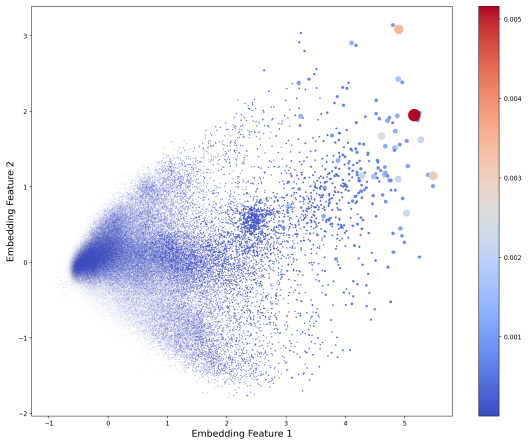

Fig. 2. Wikilinks network plotted using node2vec embeddings.

As mentioned we can see three main threads in our network, where the middle one is clearly more important according to the PageRank values. Our explanation for this was that there exist one main thread of Wikipedia which consists of very important articles and where the knowledge is based. But we all know that there are many irrelevant articles on Wikipedia and out hypothesis was that those articles are likely to be the ones spread around the main thread. To prove our hypothesis we used the information about number of views of Wikipedia pages and set threshold to cut all pages that were visited less than 5 times in the last two months. Those articles are most likely to be completely irrelevant. We again plotted the network using the same procedure and we can see the result in Figure 3. Our hypothesis was correct, we can clearly see the main thread of Slovene Wikipedia, with some of the most important pages being Slovenija, USA, English Language, etc. According to PageRank, the most important page is VIAF (The Virtual International Authority File), which is interesting, but in a way logical since many important pages link to VIAF and therefore receives high PageRank.

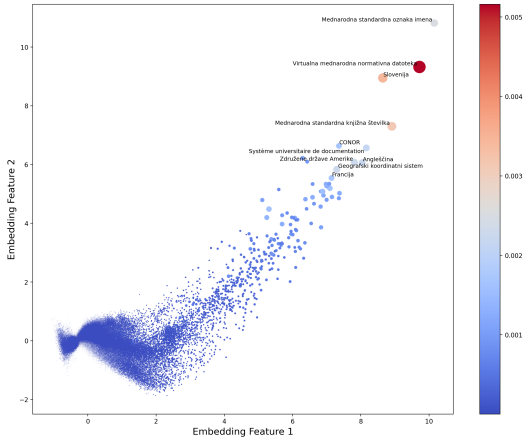

Fig. 3. Wikilinks network with removed unvisited nodes plotted using node2vec embeddings.

B. Predictive analysis. The fist problem that we were tackling was **link prediction**. In the sense of Wikipedia, this means that we want to be able to recommend which articles writer of certain page should include in the page. This is extremely important for the community, since we want to have encyclopedia that is as informative and homogeneous as possible.

Again, the idea was to encode the nodes into vector embeddings using node2vec algorithm, which we will use to train the classifier which will serve as our prediction/recommendation model. But since node2vec algorithm encodes information the neighbourhood of the nodes using random walks, we must split of data into train and test datasets before running node2vec. Therefore we randomly removed 20% of the edges and store them as positive cases in our test dataset and take the same amount of non-edges as negative cases for our test dataset. Likewise we performed also for train dataset, were we took the remainder of the edges (80%) and the same amount of non-edges. Next, we performed node2vec on the reduced network to obtain the node embeddings. And since we want to predict the edges, construct our dataset such that we concatenate the embedding of nodes between the link exist (considering also the link direction). The final result were dataset consisting of more than 18 million rows which we used to train our classifier. But since the size of our dataset is such that training is computationally impossible, we had to perform **random**

sampling to train our model in batches. For each batch we sampled 5% of the edges and train the model. The final result of the accuracy of our model can be seen in 2.

Table 2. Table describing data or methods.

	AUC	Accuracy	Recall	Precision
CNN	93.6	93.7	91.9	95.3

For our prediction model we constructed **convolutional neural network** upon which we trained our data. If we first focus on the results, we can see that we created the model which is extremely accurate, with accuracy 93.7. This means that we are able to efficiently predict whether two pages should be connected or not. The real world application of our model would be helping the writer link the right pages by recommending them in on the fly. For implementation we used TensorFlow library. The architecture of the model was relatively simple but efficient, consisting of one convolutional layer (with 128 filters of kernel size 5) and two dense layers (the first of size 32 and with ReLU activation function and the second with sigmoid activation function). For model compilation we used binary crossentropy and Adam optimizer.

Our second problem that we tackled was **community detection**. As mentioned, categories in Wikipedia are not well structured since everyone can create their own category. We discovered that more than $20k$ categories of total **66.288** contains only one article, meaning that they are totally useless. In addition, the categories on Wikipedia are not hierarchical, which again give us a lot of confusion. Our goal was to find 5-10 meaningful categories in which the articles would be sorted.

The first approach to solve this problem was to use classical network science algorithms for community detection (Louvain, Infomap, SBM, etc.), but since our network is so large, it would be computationally impossible to solve this using their pure implementations. We ran fast label propagation but the result we got was more or less useless since the majority of pages were in a single category. One approach would be to split our network into sub-graphs and perform also other community detection algorithms, but we had an idea to tackle this using the **lexical similarity** between the articles.

Therefore we constructed new network (based on the lexical similarity), where the links represent the connection between similar articles. For the similarity measure we used **TF-IDF** on the first paragraph of each page. Lastly, we needed to specify the threshold which tells us whether to link two pages or not. We set the threshold such that we obtained roughly the same as the original network.

Next, to find the communities we wanted to again convert the network into vector space and in the vector representation perform clustering using some of unsupervised machine learning approaches. To do that we again performed **node2vec** algorithm on the newly constructed network and plot it in two-dimensional space. We could clearly see 5 clusters in the data, therefore we ran **K-Means** algorithm to obtain them.

But as we can see in Figure 4, we couldn't exactly obtain the communities we expected. We can clearly see that that the majority of articles are in the center of the plot (we classified them as Slovenia related), whereas periphery mainly consist of articles that are related to foreign topics. The reason for the symmetry was probably thresholding. Our conclusion was that community detection in our network is very hard problem and we would need to use much more granularity to obtain the clusters.

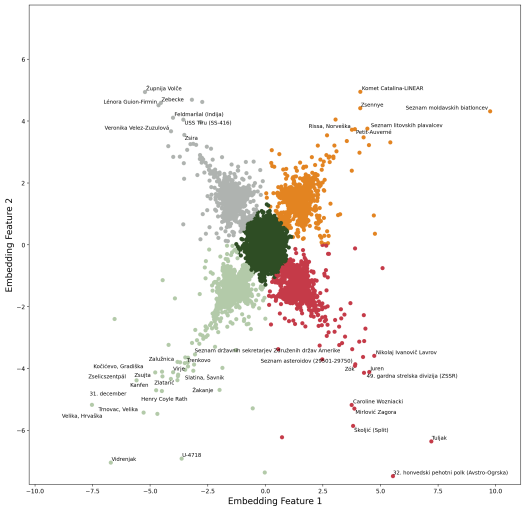

Fig. 4. Communities in Wikilinks (lexical) network using node2vec and K-Means.

Discussion

The main focus of our project was to create a model for recommending editors/volunteer on Wikipedia the links to include in their article. The problem we were trying to solve is that no one can possibly be aware of all the articles existing on Wikipedia, and with that many mistakenly or forgotten references exists in the articles. With our model we are able to efficiently and accurately predict the link one must include in their article based on the link he/she already included. With that we solve the problem of homogeneity of Wikipedia and help the community build the encyclopedia that is as informative as possible. The project also consist of different interesting finding we found during the network analysis and try to present them in a meaningful way. The future work would be of course implementing our model into a working solution which would be presenting the recommendations in a user friendly way which would be updated regularly on the fly.

1. V. Miz, K. Benzi, B. Ricaud, and P. Vandergheynst. Wikipedia graph mining: dynamic structure of collective memory. *arXiv:1710.00398 [cs]*, February 2018. URL http://arxiv.org/abs/1710.00398. arXiv: 1710.00398.
2. B. de La Robertie, Y. Pitarch, and O. Teste. Measuring Article Quality in Wikipedia using the Collaboration Network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 464–471, Paris France, August 2015. ACM. ISBN 978-1-4503-3854-7. . URL https://dl.acm.org/doi/10.1145/2808797.2808895.
3. B. Wei, J. Liu, Q. Zheng, W. Zhang, C. Wang, and B. Wu. DF-Miner: Domain-specific facet mining by leveraging the hyperlink structure of Wikipedia. *Knowledge-Based Systems*, 77: 80–91, March 2015. ISSN 0950-7051. . URL https://www.sciencedirect.com/science/article/pii/S0950705115000088.
4. K. Baraniak, M. Sydow, J. Szejda, and D. Czerniawska. Studying the Role of Diversity in Open Collaboration Network: Experiments on Wikipedia. In A. Wierzbicki, U. Brandes, F. Schweitzer, and D. Pedreschi, editors, *Advances in Network Science*, Lecture Notes in Computer Science, pages 97–110, Cham, 2016. Springer International Publishing. ISBN 978-3-319-28361-6. .
5. Zecheng Zhang, Yuan Shi, and Xinwei He. Evolution and Link Prediction of the Wikipedia Network. page 9, 2019.
6. Armand Boschin and Thomas Bonald. Enriching wikidata with semantified wikipedia hyperlinks. In *Wikidata@ISWC*, 2021.
7. L. Martiniano. Complex Network Analysis: Wikipedia Map of Science, December 2020. URL https://medium.com/swlh/complex-network-analysis-wikipedia-map-of-science-a35a3a23e453.