Topological Approach to Wikipedia Page Recommendation

Maksym Opirskyi[1] and Petro Sarkanych[2,3,4]

[1] Ukrainian Catholic University, Lviv, Ukraine
opirskyi@ucu.edu.ua
[2] Institute for Condensed Matter Physics, Lviv, Ukraine
[3] Centre for Fluid and Complex Systems, Coventry University, Coventry CV1 5FB, United Kingdom
[4] L4 Collaboration & Doctoral College for the Statistical Physics of Complex Systems, Leipzig-Lorraine-Lviv-Coventry, Europe
sarkanyp@coventry.ac.uk

Abstract. Human navigation in information spaces has increasing importance in ever-growing data sources we possess. Therefore, an efficient navigation strategy would give a huge benefit to the satisfaction of human information needs. Often, the search space can be understood as a network and navigation can be seen as a walk on this network. Previous studies have shown that despite not knowing the global network structure people tend to be efficient at finding what they need. This is usually explained by the fact that people possess some background knowledge. In this work, we explore an adapted version of the network consisting of Wikipedia pages and links between them as well as human trails on it. The goal of our research is to find a procedure to label articles that are similar to a given one. Among others, this would lay a foundation for a recommender system for Wikipedia editors, which will suggest links from the given page to the related articles. Our work is therefore providing a basement for enhancing the Wikipedia navigation process making it more user-friendly.

Keywords: recommender system · information network · random walk

1 Introduction

Over the last five years, the total number of web pages has nearly doubled from 900 million pages in 2014 to more than 1.7 billion pages in 2019 [1]. With this increase in size, it is obvious that the amount of information has grown as well, making its retrieval much harder.

The Web itself is a good example of a large information network, where pages play role of nodes and links between pages are the edges. Furthermore, information networks can often be divided into smaller pieces, each with narrower specialization. Even then, the task of finding a particular bit of knowledge is not trivial, since smaller networks

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 89–97

can have complex structure as well. However, having the goal, humans can search for it quite efficiently, as shown by West and Leskovec [2]. This fact is explained in a way, that often the structure of the network preserves real-world properties, in a sense that two concepts that are related have higher chance to be connected (or to be within the small range of each other) in the network. Even though humans do not know the global structure of connections between common concepts, they usually have an idea or vague understanding of their relatedness and can efficiently exploit this background knowledge when searching for information [3, 4].

Nevertheless, the notion of the relatedness can be ambiguous and often is context-dependent. For example, consider concepts Water, Molecule, and Swimming. Clearly, Water and Molecule are related closely, as well as Water and Swimming. However, if one is, say, looking for information about the water states, she barely needs anything related to swimming. Moreover, the structure of the network does not only rely on the semantic relatedness between the concepts it describes. The structure can be dictated by the purpose of the web resource and its design. These arguments demonstrate the need for the tools that would help humans to navigate through the networks efficiently.

In this work, we pursue the goal of developing a procedure that could mark similar pages based on the topological properties of the network. To this end, we explore the subset of the Wikipedia network. In this network, articles are represented by nodes and links between articles are edges in the corresponding graph. Wikipedia is a good approximation to other information networks as it is a collection of concepts and its structure partially represents human knowledge. On the one hand, this would allow unveil the connection between the concepts. On the other hand, one can use this method for purely applied tasks. One such task is the recommendation of similar pages for the ”See also” section of an article. This section is either missing or is manually constructed by the Wikipedia editor.

Implementation of such a procedure would allow enhance the navigation by proposing articles that are relevant for the user. Similarly, editors would be able to use it to create ”See Also” sections. This would make Wikipedia browsing experience more pleasant for both the information seekers and the contributors.

The rest of the paper is structured as follows. In Section 2, we give an overview of the related work. The primary dataset that we are going to use and its structure is described in Section 3. Section 4 presents the vision of the approach and our plans for future work.

2 Related Work

Our research is related to three separate topics. First is the navigation in information networks. Unlike in Euclidean space, finding the shortest path in a network is not as trivial as drawing a straight line. Second is the study of patterns in search strategies. This might allow analyze semantic connections between concepts. Last but not the least, is the development of recommender systems that are known to provide users with relevant content. Below we will give short reviews of important papers in each of the areas separately.

2.1 Navigation in Networks

Taking into account that the topology of a network possesses neither short nor long ordering, the structure of the network far away from an element does not allow predict its position. Thus, developing efficient navigation algorithms for network search is not a trivial task. In 2000, Kleinberg showed that networks that possess small-world property could be efficiently navigable [4]. In particular, he proved that time needed by any decentralized algorithm is bounded by polynomial in $log N$, where N is the number of nodes in the network, provided that network is constructed using a certain model.

This finding was used in [5] to perform searches on social networks. The resulting search paths were similar to human-like ones, and the efficiency depended on the background knowledge of the network structure.

Further, Trattner et al. [7] applied decentralized search algorithms with the so-called hierarchical knowledge of the network to model human searches in information networks. They use an algorithm for decentralized search, which is based on previous work and utilizes given hierarchical knowledge. Evaluation of their models of hierarchies is done by comparing them to human navigation paths. They conclude that the model using a hierarchy based on the network topology produces results that are the closest to human search trails.

In [6], West et al. use data gathered from Wikispeedia [20], an on-line game, where players are given a source node and a target node, and the goal is to reach the target using the least number of steps, clicking on the hyperlinks exclusively. The dataset consists of players' game paths. Authors use this information to develop a new information-theoretic metric, which measures semantic distance between concepts that Wikipedia articles represent. They also develop an approach to filter out concepts that have a small distance to the goal concept but are, in fact, irrelevant. They use concepts that are labelled as relevant or irrelevant by humans to teach a neural network to do this kind of filtering. Proposed method outperforms Latent Semantic Analysis, which is validated by the psychological community. However, one flaw of their approach is the inability to calculate the similarity if one of the articles never occurred in any trail. In order for the method to work well, a huge amount of paths to calculate probabilities is required.

In light of the previously mentioned studies, Niebler et al. [8] investigate the ability to extract useful semantic information from various web resources. They argue the usability of game-like navigational data like Wikispeedia in computing semantic relatedness between concepts that Wikipedia contains. In turn, they study unconstrained data, where users are allowed to jump to any other article (i.e. use search field) and try to see whether it is suitable for extracting semantic information. Using adaptation of their previous work, they show that unconstrained navigation data indeed can be used for semantic extraction.

West and Leskovec [9] further explore human navigation, this time comparing it to automatic navigation. The main requirement to the agents is to be local, in a sense that every step in a search can only be based on the local network characteristics and a given target. They used two types of synthetic agents. First group of agents only calculates

properties of every possible node before making a step. Second group contains algorithms that learn either from the user paths or from previous steps. The authors show that agents can perform the task of transition from goal to target more effectively than humans. They conclude that the background knowledge humans possess in not necessary for successful Wikipedia navigation. However, humans, in contrast to algorithms, rarely get entirely lost. It is explained by the fact that humans can make longstanding plans.

In [10] Lamprecht et al. study how the structure of each article influences user navigation. They firstly confirm that lead sections and infoboxes[1] of articles contain links to more general concepts. Using data about user click behavior in Wikipedia as well as Wikispeedia game paths, they explore what influences users' trajectories. Then, they compare distributions of the ground truth game clicks and the clicks suggested by the first-order Markov model with different choices of transition probabilities to examine the influence of factors like the structure of the page or its generality on human choice. Next, they analyze only goal-oriented navigation using the same approach but construct the models for each step of the path of each length and every optimal path length. Overall, they confirm that human navigation is biased towards the article structure.

2.2 Patterns in Search Strategies

In [2] West and Leskovec ask why human navigation paths while different from optimal ones, are still efficient on average. They confirm that both high degree and high similarity are valuable for the choices, but at the early stages of the game people choose high degree nodes and then the textual similarity of the articles becomes significant while approaching the end of the game. Furthermore, the similarity of articles is more significant for successful games. Finally, using the Markov model with two types of click probability - binomial logistic and learning-to-rank (multinomial) - authors develop a method to predict the target of the game having only the beginning of the path. This method beats the baseline that predicts the target by choosing an article that has the highest textual similarity with the last article in a given path.

While successful games contain a considerable amount of useful information, unfinished game paths can provide valuable findings as well. Scaria et al. [11] aim to identify search abandonment reasons by analyzing unfinished paths of Wikispeedia players and comparing them to finished ones. They report that properties like PageRank or indegree of the target are influential on game abandonment since the difference between these properties in finished and unfinished games is statistically significant. On the other hand, source properties are not a factor for abandonment, although the difference in the latter case is statistically significant as well. Another finding is that unsuccessful missions can be characterized not as "getting lost" in the network, which can be described as increasing shortest path length (SPL) to target or source, but as orbiting close around the target and inability to find a required link. Authors also analyze back clicking patterns, reporting that users have a higher probability of backtracking when SPL or *tf-idf* distance [22] (calculated as one minus *tf-idf* similarity) to the target is getting bigger.

[1] Infobox is the short summary of an article, usually placed in the top right corner.

What is more, backtracking is more probable in these situations for better players, meaning that they not only can better plan their moves but also better understand when they are getting lost. Authors also develop a model that predicts whether a user will abandon the mission, whether the next click will be back-click and whether a user will give up the search after current node, having the first couple of clicks as input.

Contrasting to [2] research conducted in [12] focuses on strategies in a scenario where no explicit target is specified. Rodi et al. explore Wikipedia user click data and consider the last article visited as the target. To find patterns of search strategies, authors develop a vector representation of the page where coordinates represent 13 abstract topics. Authors simulate human paths with random walks with transition probabilities based on click fractions available from the dataset and compare them to Wikispeedia real paths. They report that Wikipedia readers tend to start from abstract pages and narrow down to the specific ones, while Wikispeedia players first head to hubs and then narrow down. While semantic distance between consecutive pages is changing the most at the beginning and the end of the path, distance to the target keeps decreasing monotonically.

2.3 Recommender Systems

Algorithms beneath the recommender systems are traditionally classified into content-based and collaborative [15, 14]. Drawbacks of each type of algorithms alone can be overcome by combining them, which results in hybrid recommendation techniques [15]. However, there are other approaches. For example, in [16] collaborative filtering algorithm is used. Its central problems - cold start and data sparsity - are attacked by using Wikipedia for extracting similarity information between items and using it to compute artificial ratings for items that are not rated by the user. Authors use textual similarity of the articles, as well as category similarity and degree information. This differs from our approach, as we aim to extend information used by taking into account the topological properties of the network and users' trails.

In [17] a bot that routes tasks for Wikipedia editors is developed. It uses the textual information and link structure available from Wikipedia to make recommendations about work that the editor should do based on the history of their previous contributions.

Recommender system described in [13] utilizes a hybrid approach to combine different sources that can be used to produce music recommendations. It uses Wikipedia to get information that is most relevant to the user's profile. In this setting, data from Wikipedia is just a supportive mechanism to the whole system.

Hickcox and Min in [18] analyze Wikispeedia network to develop a recommender system for similar Wikipedia articles. To this end, they set up a random walker with a step probability dependent solely on the next node properties. The nodes that are occurring the most number of times in a series of random walks are considered as similar. As a recommendation quality measure, they used the *tf-idf* similarity. Five (including uniform random walk) out of seven probability schemes scored nearly the same value. Taking into account that for the uniform random walk distance from the source grows

as a power law $R \propto t^{v}$, where v is so-called Flory exponent (see, e.g. [21])[2], this leads to the conclusion that the methods suggested in [18] perform poorly, and, most probably, recommend nearest neighbors of a given article. To the best of our knowledge, [18] is the only work, where tasks similar to ours were put forth. We describe how our approach differs from one considered by Hickcox and Min in Section 4.

3 Data

In the research, we use the data available from human-computation game Wikispeedia. The game can be understood as a walk on the network, where articles are represented by nodes and directed link between nodes exists if there is a hyperlink from one article to another. For now, we consider two datasets: Wikispeedia graph and collection of human game paths.

We choose this dataset for two reasons. Firstly, its size allows for easier computations and faster testing of hypotheses. Secondly, it still represents an information network, since it is a version of Wikipedia for schools from 2007.

Wikispeedia graph contains 4 592 nodes and 119 882 links. We computed the basic properties of this network to get initial knowledge about its structure. They are summarized in Table 1.

Table 1. Basic properties of Wikispeedia network. $\langle l \rangle$ denotes average shortest path length, r is assortativity by degree, $max_{(u,v)}d(u,v)$ is network diameter, $\langle c \rangle$ is average local clustering coefficient, C is global clustering coefficient, $\langle C_B(v) \rangle$ is average betweenness centrality.

Property	**Value**
$min_{(v \in V)} deg^{-}v$	0
$max_{(v \in V)} deg^{-}v$	1551
$min_{(v \in V)} deg^{+}v$	0
$max_{(v \in V)} deg^{+}v$	294
$\langle deg^{-}v \rangle$	26.11
$\langle deg^{+}v \rangle$	26.11
$\langle l \rangle$	3.2
r	-0.056
$max_{(u,v)}d(u,v)$	9
GCC size	4051 or 88.0%
$\langle c \rangle$	0.11
C	0.1
$\langle C_B(v) \rangle$	8915.72

Values in the table reveal some interesting properties of the network. Firstly, this network is disconnected, though a significant fraction of nodes forms a connected component. Secondly, Wikipedia is a typical small-world network. We can observe large

[2] This exponent is dependent on the properties of space.

values of maximum indegree and outdegree, which corresponds to hubs – the pages that point to or are pointed to by many other pages. Furthermore, the average shortest path is only 3.2, which means that one needs three clicks on average in order to get to any page in the network. Empirical evidence of the small world property can be found in Fig. 1: degree distributions resemble power-law function.

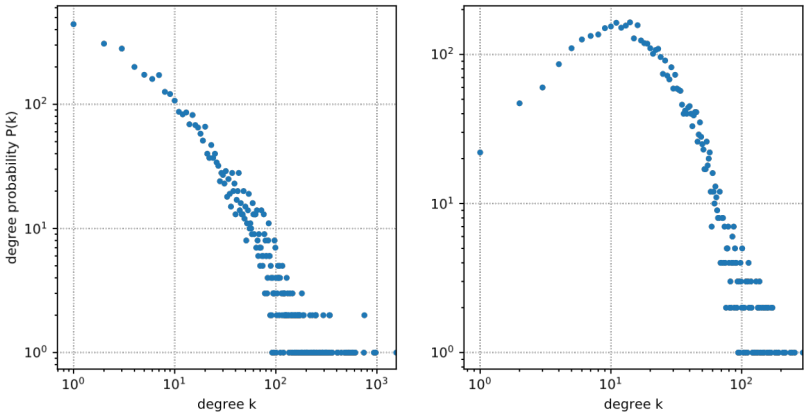

Fig. 1. Degree distributions of Wikipedia graph: (left) indegree distribution, (right) outdegree distribution

Game paths data are organized as a file where each row contains a sequence of article titles, representing a single path. In the game setting, the user is allowed to return to pages that she visited earlier. Back clicking is denoted by ”<” symbol. There are 51 318 finished and 24 875 unfinished paths in the dataset.

4 Approach

In a number of previous works, human navigation was modelled by Markov chains. Singer et al. in [19] successfully applied the Markov model without memory to model human page-level navigation. Thus, our first approach is to run a random walk on a Wikipedia graph. When modelling human trails on the network, loops need to be discarded from the consideration, because according to the game rules, these can just be realized by clicking the ”Back” button. Therefore, for this purpose, self-avoiding walks (walks where each node can be visited only once) have to be considered.

As we have already mentioned in Section 2.3, attempts to mark similar pages of Wikipedia have already been made, but with questionable success. Since one of our primary tasks is to develop a recommender system for the ”See also” section of a Wikipedia page, then we are talking about adding ”missing” edges to the network. With this interpretation, limiting ourselves to steps only between the neighboring nodes does not fully fit the goal, since additional links can only be added between the nodes on the distance at least 2. Therefore, our walker should be allowed to make jumps of longer

distances. If right transition probabilities are selected, we will be able to find out which regions of the network are the most visited. These regions could serve as a cluster from which we could obtain recommendations for an initial article. There is a number of choices of transition probabilities we would like to test:

- $p_{ij} \propto \frac{sim(tf-idf(i),tf-idf(j))}{d(i,j)^{\alpha}}$, where $sim(i,j)$ is cosine similarity between vectors i and j, and α is a positive parameter;
- $p_{ij} \propto r_{ij}$, where r_{ij} is a Pearson similarity of nodes i and j. Taking into account the definition of r_{ij}, in this case only steps of maximum length of 2 can be made;
- $p_{ij} \propto k_{ij}$, where k_{ij} equals to the number of transitions from node i to j through all the game paths.

In all options above, probabilities need to be properly scaled, to sum up to one for each node. For now, we omit normalizing coefficients for the sake of simplicity. In the first two cases, only topological properties of the Wikispeedia network are considered, while in the third scenario, human experience and expectations are indirectly taken into account.

5 Conclusion

In this work, we presented a vision of a random walk based recommender system for obtaining Wikipedia articles that are similar to the given one. To this end, we explored topological properties of Wikispeedia as well as its user navigation history, since we want our random walkers to incorporate this knowledge. We proposed a number of walkers with different transition probabilities that should grasp the similarities between articles. Apart from that, we performed a literature survey on the topics related to our work, discussed the problems in network navigation and approaches for implementing recommender systems. We believe that random walk methods can be applied in the context of recommender systems and that the navigation history can enhance their performance.

References

1. Total number of websites. https://www.internetlivestats.com/total-number-of-websites/
2. West, R., Leskovec, J.: Human wayfinding in information networks. In: 21st International Conference on World Wide Web, pp. 619–628. ACM (2012)
3. Kleinberg, J.M.: Navigation in a small world. Nature **406**, 845 (2000). doi: 10.1038/35022643
4. Kleinberg, J.M.: Small-world phenomena and the dynamics of information. In: Advances in neural information processing systems, pp. 431–438. MIT Press (2002)
5. Adamic, L., Adar, E.: How to search a social network. Social networks **27**(3), 187–203 (2005)

6. West, R., Pineau, J., Precup, D.: Wikispeedia: an online game for inferring semantic distances between concepts. In: 21st International Joint Conference on Artificial Intelligence (2009)
7. Trattner, C., Singer, P., Helic, D., Strohmaier, M.: Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In: 12th International Conference on Knowledge Management and Knowledge Technologies, article No. 14 pp. 1–8. ACM (2012)
8. Niebler, T., Schlör, D., Becker, M., Hotho, A.: Extracting semantics from unconstrained navigation on Wikipedia. Künstliche Intelligenz **30**(2), 163–168 (2016)
9. West, R., Leskovec, J.: Automatic versus human navigation in information networks. In: 6th AAAI International Conference on Weblogs and Social Media, pp. 362–369. AAAI (2012)
10. Lamprecht, D., Lerman, K., Helic, D., Strohmaier, M.: How the structure of Wikipedia articles influences user navigation. New Review of Hypermedia and Multimedia **23**(1), 29–50 (2017)
11. Scaria, A.T., Philip, R.M., West, R., Leskovec, J.: The last click: why users give up information network navigation. In: 7th ACM International Conference on Web Search and Data Mining, pp. 213–222. ACM (2014)
12. Rodi, G.C., Loreto, V., Tria, F.: Search strategies of Wikipedia readers. PLOS ONE, **12**(2), p.e0170746 (2017). doi: 10.1371/journal.pone.0170746
13. Bostandjiev, S., O'Donovan, J., Höllerer, T.: TasteWeights: a visual interactive hybrid recommender system. In: 6th ACM Conference on Recommender Systems, pp. 35–42. ACM. (2012)
14. Lops, P., de Gemmis, M.. Semeraro, G.: Content-based recommender systems: state of the art and trends. In: Ricci F., Rokach L., Shapira B., Kantor P. (eds.) Recommender Systems Handbook, pp. 73–105. Springer, Boston, MA (2011)
15. Asanov, D.: Algorithms and methods in recommender systems. Berlin Institute of Technology, Berlin, Germany (2011)
16. Katz, G., Ofek, N., Shapira, B., Rokach, L., Shani, G.: Using Wikipedia to boost collaborative filtering techniques. In: 5th ACM Conference on Recommender Systems. pp. 285–288. ACM (2011)
17. Cosley, D., Frankowski, D., Terveen, L., Riedl, J.: SuggestBot: using intelligent task routing to help people find work in Wikipedia. In: 12th International Conference on Intelligent User Interfaces, pp. 32–41. ACM (2007)
18. Hickcox, J., Min, C.: Customized Random Walk for Generating Wikipedia Article Recommendations. http://snap.stanford.edu/class/cs224w-2015/projects_2015/Customized_Random_Walk_for_Generating_Wikipedia_Article_Recommendations.pdf
19. Singer, P., Helic, D., Taraghi, B., Strohmaier, M.: Detecting memory and structure in human navigation patterns using Markov chain models of varying order. PLOS ONE **9**(7), p.e102070 (2014)
20. Wikispeedia. https://www.cs.mcgill.ca/~rwest/wikispeedia/
21. Tamm, M.V., Polovnikov, K.: Dynamics of polymers: classic results and recent developments. arXiv preprint arXiv:1707.09885 (2017)
22. Manning, C.D., Raghavan, P., Schüutze, H.: Introduction to Information Retrieval Cambridge University Press, Cambridge (2008)