Readers' Demanded Hyperlink Prediction in Wikipedia

Laxmi Amulya Gundala and Francesca Spezzano
Computer Science Department
Boise State University, USA
{laxmiamulyagundala@u.,francescaspezzano@}boisestate.edu

ABSTRACT

In this paper, we describe our on-going research on the problem of predicting needed hyperlinks between pairs of Wikipedia pages (u, v) that are not connected, yet show readers' search navigation from u to v. We propose a solution that first estimates how long will these searches last and then predicts new hyperlinks according to descending order of duration. Our initial experimental results show that our best solution achieves an AUROC of 0.77 on the Wikipedia Clickstream dataset and a precision@20% of 1.0 and significantly beats the baselines.

CCS CONCEPTS

• **Information systems** → **Wikis**; **Data mining**; *Web log analysis*;

KEYWORDS

Hyperlink prediction; Estimating event duration; Wikipedia Clickstream.

ACM Reference Format:
Laxmi Amulya Gundala and Francesca Spezzano. 2018. Readers' Demanded Hyperlink Prediction in Wikipedia. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3184558.3191644

1 INTRODUCTION

Wikipedia is the online encyclopedia that anyone can edit. Many editors contribute every day to expand Wikipedia content and keep it updated. Also, many users read Wikipedia every day [5]. Reading and navigating Wikipedia is facilitated by a hyperlink network that connects related articles.

However, many hyperlinks are currently missing. When readers do not find the link to a page v from their current page u, they use the search box (on the top right corner of page u) to navigate from page u to page v. Some of these searches can be casual or because of an ongoing trend in news or social media. For instance in Wikipedia, during February 2016, the time when Donald Trump was nominated as Republican Nominee, users navigated from his Wikipedia page to different other Wikipedia pages like *Trump University*, *Hollywood Walk of Fame*, *Hillary Clinton*, etc. Though the number of searches between those pages was in thousands, they were casual as they didn't continue after that period.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191644

Other searches, instead, suggest the demand of a physical link from page u to page v to improve users' navigation. For instance, in Wikipedia, there were many searches between two pages *Doctor Strange (film)* and *Baron Mordo* in February 2016 and users continued to navigate between these two pages even until April 2016. Later on, a hyperlink from *Doctor Strange (film)* to *Baron Mordo* was created after April 2016.

In this paper, we describe our on-going research on the problem of predicting needed hyperlinks between pairs of non-connected pages (u, v) that show readers' search propagation from u to v. Our proposed solution leverages the information on the estimated duration of these searches and predicts new hyperlinks according to descending order of duration. In fact, the longer readers keep searching from u to v, the higher the probability that a hyperlink from u to v is needed to improve readers' navigation. Our initial experimental results show that our best solution achieves an AUROC of 0.77 on the Wikipedia Clickstream dataset and a precision@20% of 1.0 and significantly beats the baselines with 9% of AUROC improvement.

2 WIKIPEDIA CLICKSTREAM DATASET

Wikipedia Clickstream is a Wikimedia's research project [1] in progress. The dataset [10] consists of pairs (referrer page, resource page) obtained from the extracted request logs of Wikipedia. There are eight (non-contiguous) months' datasets released till December 2017, starting from January 2015. Each dataset contains a set of tuples of the from *(prev, curr, type, n)* where:

prev is the referrer URL or Wikipedia page title if the referrer is within Wikipedia.

curr is the title of the webpage the client requested or Wikipedia page title if the resource page is within Wikipedia.

type describes *(prev, curr)* as (a) *link* if the referrer and request are both Wikipedia pages and the referrer links to the request, (b) *external* if the referrer host is not en.wikipedia.org, and (c) *other* if the referrer and request are both Wikipedia pages but the referrer does not link to the request. This can happen when clients search or spoof their referrer.

n is the number of occurrences (greater than 10) of the (referrer, resource) pair. Considered as number of hits from *prev* to *curr*.

In this paper, we focus on pairs of pages having *type='other'*. They refer to pairs of pages from Wikipedia which do not have a direct link between them, but users navigated through search bar of the Wikipedia pages.

For our initial set of experiments, we consider February 2016 (27M of tuples), March 2016 (25M of tuples), and April 2016 (21M of tuples) datasets as they are the longest consecutive months available in Clickstream. For new links ground truth we considered August

[1] https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream

Our Framework	AUROC
Basic Approach (RBF SVM)	0.500
Fine-grained Approach	
Survival Analysis (AFT-LogNormal)	0.528
Regression (Lasso)	**0.769**
Best Baselines	AUROC
Node2Vec	0.583
Classical Approach (K-NN)	0.679

Table 1: Comparison of AUROC values for hyperlink prediction between our solution and the best baselines. The table reports the values for the best models only.

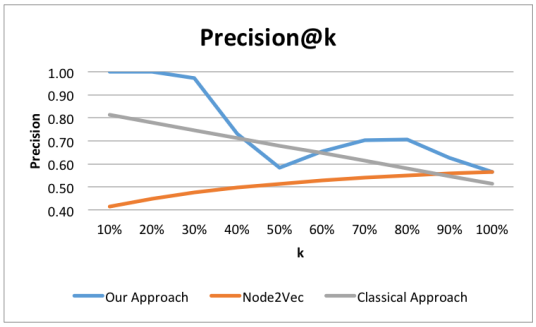

Figure 1: Precision@k curve.

2016 dataset (24M of tuples, out of which 14M have *type='link'*). Also, we filtered out from February 2016 *other* types of tuples having *prev* or *curr* equal to the Wikipedia Main Page. This is because the Main page changes every day and searches from or to this page represent noise in the dataset.

3 RELATED WORK

There is plenty of work on the hyperlink prediction problem in Wikipedia. Adafre et al. [1] proposed an approach to find missing hyperlinks in the network by considering a Wikipedia corpus and its underlying abstract words. They stated that similar pages should have similar hyperlinks. Noraset et al. [3] did a similar work by considering the text in Wikipedia pages. West et al. [9] proposed a method based on principal component analysis of the hyperlink adjacency matrix. In another work, West et al. [7] used human navigation logs available from The Wiki Game[2] and Wikispeedia [8] to identify missing links in Wikipedia. Recently, Paranjape et al. [4] addressed the problem of suggesting hyperlinks to add in the encyclopedia that improve readers' navigation and maintain a high-quality link structure. In some sense, they are close to our work because they are suggesting links that are useful in the future. However, their framework is more general than ours. Also, there is a main difference between our work and the one of Paranjape et al. when we have to predict the link between two pages $(prev, curr)$ having $type =$"$other$". More specifically, their method relies on the number of hits from $prev$ to $curr$ only, while we consider the estimated duration, computed by using many other network-based variables, of how long readers will keep searching from $prev$ to $curr$. Moreover, the distribution of hits for pairs of pages having $type =$"$other$" in February 2016 dataset is the same for pairs of pages that will have or not a hyperlink, suggesting that the number of hits alone is not a good predictor for our problem.

4 HYPERLINK PREDICTION FRAMEWORK

In this paper, we address the problem of predicting hyperlinks between pairs of pages (u, v) of type "*other*" in the Clickstream dataset. If readers persistently keep searching for page v from page u, then this is the signal that a hyperlink may be needed from u to v to improve their navigation.

[2]https://thewikigame.com

We propose a *novel* approach to solve the problem in two steps. First, we focus on the problem of predicting how long will readers keep searching from page u to page v. Second, once the duration is estimated, we predict new hyperlinks according to their descending order.

We performed a two-sample one-tail t-test to test our hypothesis that the longer readers keep searching from page u to page v, the higher is the probability that a hyperlink will be added from u to v. The p-value for a significant level $\alpha = 0.01$ resulted less than 0.001 on the Clickstream dataset considered, suggesting that there is strong statistical evidence to validate our hypothesis. More specifically, readers keep searching for 2.02 months for pairs of pages that end up to be linked vs. 1.56 months for pages that will not be linked.

We propose two supervised learning approaches for the problem of predicting the duration of the searches. Given a set of features, the *basic* approach consists of learning a binary classifier to predict whether or not the searches will last in the future. The second and more *fine-grained* approach consists of estimating how long the searches will last by modeling the problem via survival analysis [6] or as a regression task.

We used the following set of features in our preliminary evaluation: node degree, number of hits from page u to page v, common neighbors, Jaccard similarity, Adamic-Adar similarity, preferential attachment score, local clustering coefficient, Pagerank of both pages u and v, and the Hadamard (or entrywise) product of Node2Vec feature vectors $F(u)$ and $F(v)$ [2].

5 INITIAL RESULTS

We divided the data into three time periods: we consider candidate pairs of pages for link prediction those appearing as type "*other*" in February 2016; we trained the model to estimate the searches duration in March and April 2016, and tested hyperlink prediction in August 2016. We used majority under-sampling to deal with class imbalance.

Our results show that we achieve an AUROC of 0.78 (with random forest) on the problem of predicting whether or not the searches from page u to page v will last (basic approach), while regression performs better than survival analysis (fine-grained approach) on the problem of predicting the duration of these searches

(Mean Absol. Error of 0.094 vs. 0.118 and Pearson Corr. Coeff. of 0.91 vs. 0.32).

Table 1 (top rows) shows how our proposed framework performs on hyperlink prediction. In this experiment, we first trained the model (classification, survival analysis, or regression) to predict how long (or "if" in the case of classification) the searches will last. Then we computed the AUROC between these predicted times and the class values (1 if the hyperlink has been created later on, 0 otherwise). As we can see, using the information on whether or not the searches will last is not sufficient to predict new hyperlinks (AUROC of 0.5 with SVM with RBF kernel). Instead, if we are using a more fine-grained approach with Lasso regression that is predicting how long the interaction will last, we can achieve an AUROC of 0.769.

To compare our results, we considered hits, Jaccard Similarity, Adamic-Adar, preferential attachment, and cosine similarity between Node2Vec feature vectors $F(u)$ and $F(v)$, individually as baselines and assumed their value to be an approximation of the duration of the searches from page u to v. As we can see in Table 1 (bottom rows) our approach outperforms Node2Vec, the best performing baseline, that achieves an AUROC of 0.583. Also, our result is better than a classical link prediction approach that considers all the features in input to a classifier, do not consider search duration, and can predict new hyperlinks with an AUROC of 0.679.

We also computed the precision@k curves for our best approach, the best baseline (Node2Vec), and the classical approach. These curves are shown in Figure 1. We see that our approach generally achieves a better precision. In particular, we have the highest precision on top-ranked pairs of pages (precision of 1.0 for both $k = 10\%$ and $k = 20\%$, and of 0.97 for $k = 30\%$). The drop at $k = 50\%$ means that no hyperlinks are found for $30\% < k \leq 50\%$, which suggests further research is needed to improve.

6 NEXT STEPS

As future work, we plan to extend the set of features considered including both page content and categories' similarities to improve our current results. In fact, we observed that, by adding category-based similarity features, we can improve the performances of the basic approach. Indeed, we can achieve an AUROC of 0.83 for the problem of predicting whether or not the searches from page u to page v will last, vs. the current value of 0.78 obtained by considering network-based features only.

Also, the Wikimedia Foundation recently announced that they started releasing Clickstream datasets on a monthly basis. Currently, November and December 2017 have been released. Thus, we plan to test our approach on a larger scale once a higher number of consecutive months are released.

REFERENCES

[1] Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering Missing Links in Wikipedia *(LinkKDD '05)*. 90–97.

[2] Aditya Grover and Jure Leskovec. 2016. Node2Vec: Scalable Feature Learning for Networks *(KDD '16)*. 855–864.

[3] Thanapon Noraset, Chandra Bhagavatula, and Doug Downey. 2014. Adding high-precision links to Wikipedia *(EMNLP '14)*. 651–656.

[4] Ashwin Paranjape, Robert West, Leila Zia, and Jure Leskovec. 2016. Improving Website Hyperlink Structure Using Server Logs *(WSDM '16)*. 615–624.

[5] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why We Read Wikipedia *(WWW'17)*. 1591–1600.

[6] Ping Wang, Yan Li, and Chandan K. Reddy. 2017. Machine Learning for Survival Analysis: A Survey. *CoRR* abs/1708.04649 (2017).

[7] Robert West, Ashwin Paranjape, and Jure Leskovec. 2015. Mining missing hyperlinks from human navigation traces: A case study of Wikipedia *(WWW '15)*. 1242–1252.

[8] Robert West, Joelle Pineau, and Doina Precup. 2009. Wikispeedia: An Online Game for Inferring Semantic Distances Between Concepts *(IJCAI'09)*. 1598–1603.

[9] Robert West, Doina Precup, and Joelle Pineau. 2009. Completing wikipedia's hyperlink structure through dimensionality reduction *(CIKM '09)*. 1097–1106.

[10] Ellery Wulczyn and Dario Taraborelli. 2015. Wikipedia Clickstream Dataset. https://figshare.com/articles/Wikipedia_Clickstream/1305770 *(Website)*.