Super Awesome Wikipedia Hops Fairness Article

Blandine Seznec blandine@seznec.com
Philip Thruesen Philip@thruesen.com
Patrick Bach Andersen Philip@thruesen.com
Jaroslav Cechak Jaroslav@cechak.com
Roel Castano Roel@castano.com

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

This paper explores the user behaviour in

I. Introduction

Wikipedia is an information network with more than 5 million articles of semi-structured information (text and multimedia). Seeking the relevant (related) information among all these articles can prove cumbersome. Currently, most Wikipedia articles provide a "See also" section with some recommendations of related articles. These recommendations are added manually by the contributors. This is, of course, time consuming and error-prone, as the human contributors' recommendations might not be complete nor reflect what readers find relevant.
By utilising the available data, which include Wikipedia Clickstream data sets, specific per second/day page view count, and derived data from article structures, we believe we could improve the "See Also" section to provide unbiased, accurate recommendations.
This has been researched multiple times but we have found most of them to be based almost completely on data from Wikispeedia or The Wiki Game, which both supply paths utilised by players to reach certain, unrelated articles. This does affect how the user navigated Wikipedia and doesn't reflect natural user behaviour.
Our method for this investigation is based on real user behaviour and Wikipedia data. Using the Learning to Rank algorithm, it is possible to weight the importance of all of the different indicators and data and recommend the most useful articles for the user. By solving this problem, we would like to get insight into the following questions.

- Is it possible to predict the user hops utilizing generalized information by Wikipedia?
- How could we improve wikipedia articles fairness by altering link arrangement?

II. Related Work

Previous work has been done with link structure in Wikipedia from which we acquired key

concepts for this article. [1] explores the issue of disambiguation and detection of possible links in external texts. In addition to the main focus in automatic cross-reference of external articles, this paper provides an understanding of certain techniques used to detect potential links in articles and the proper reference (disambiguation) of terms in wikipedia. Detecting links relies in a machine learning algorithm similar to the one applied in this article which weights in different features of the articles to provide a score to all potential links and chose the final ones. Examples of features used in this implementation include link probability, relatedness of topics, disambiguation confidence, and many others. Learning to disambiguate links on the other hand, means identifying the correct meaning, for example "crane", as a large bird or a mechanical lifting machine, depending on the context (using unambiguous concepts for example) and probability of said word.
On the other hand, [2] adds on the topic of identifying missing hyperlinks by utilising data sets of navigation paths from wikipedia-based games in which users find paths between articles. This is useful for studying shortest paths between target articles but we believe this does not address the issue of aiding users by presenting them with articles that might be interesting.
[3] comes closer to the goal of pointing out missing references by making use of server logs to weight the usefulness of links that are not yet implemented. By studying the user's paths through Wikipedia they find patterns which could be shortened with missing links.

III. Learning to Rank Algorithm

The algorithm used for choosing from the list of links is called Learning to Rank. This machine learning algorithm ranks a set of 'documents' given a query based on a set of features defined previously and provides a label (grade) to these documents. The label, or relevance of a document, can be given in multiple forms:

- Binary Label
- Multi-Level Judgement
- Pairwise Preferences

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

IV. Features

This section defines each feature used in the implementation of Learn to Rank algorithm.

.1 Link Position

Our intuition behind the Link Position feature is based on two reasons. The first one is that given the way Wikipedia articles are structured, the most general description of the article is placed in the first few paragraphs before the table of contents. This section of the Wikipedia article is called the lead [6]. As described in the Wikipedia manual of style, for many people, the lead section is the only section they will read since it summarises the entire article. Later paragraphs only dive deeper into the topics outlined in the description.
The second motivation is the way people read web pages. As explained by Jakob Nielsen [5], one of the leaders in human-computer interaction, on average, users have time to read at most 28% of the words of a website. Additionally, most attention is given to the top portion of a page and later sections are merely skimmed through. People looking for different article that is somewhat related to the one currently being read might be interested in more general concepts as they contain the searched term. As explained above, more general terms happen to be heavily abundant in the first portion of Wikipedia article.

To be able to measure this feature, we count the number of characters preceding the occurrence of the link. The text of the articles is taken from the wikipedia dumps previously described and links to other wikipedia articles are found using regular expressions. As expected, links to images, external sources, etc. are ignored. Due to the way this feature is extracted, there might be slight discrepancies in feature value and the exact number of characters preceding the links.

.2 Link Order

Similar to the previous feature, Link Order is based on the fact that a wikipedia article's initially describe the topic in general terms and the hypothesis?? that the probability of clicking a link is higher the closer it is to the first term. While the link position captures more of a distance between links and their spread, link order is a simplified version of it. It conveys less information, but in a much more straightforward manner.
This feature is also extracted from the Wikipedia dumps by counting the number of links in the article. This feature captures the position of link relative to all the other links contained in the same article. The value n means that the link is the n^{th} link in the article.

.3 Community Membership

This feature captures notion of two article being in the same community of articles. The communities are computed from the graph representation of Wikipedia $G(V, E)$, where V is a set of articles and E is set of links between them. In this scenario community is $(V?, E?) = G? \subseteq G$ such that $|E?| \geq |\{\{u, v\} \in E? | u \in V? \vee v \in V?\}|$. Communities are very implicit way of clustering articles on Wikipedia that captures emerging property of interconnected articles.

.4 Symmetric Linking

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

.5 HITS and PageRanks

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

V. Results

Table 1: *Example table*

Name		
First name	Last Name	Grade
John	Doe	7.5
Richard	Miles	2

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras

nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$e = mc^2 \tag{1}$$

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

VI. Discussion

I. Subsection One

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

II. Subsection Two

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

References

[1] D. Milne and I. H. Witten, "Learning to Link with Wikipedia" *Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM '08 (2008): Web.*

[2] R. West, A. Paranjape, and J. Leskovec, "Mining Missing Hyperlinks from Human Navigation Traces: A Case Study of Wikipedia" *Proceedings of the 24th International Conference on World Wide Web - WWW '15 (2015): Web*

[3] A. Paranjape, R. West, L. Zia, and J. Leskovec, "Improving Website Hyperlink Structure Using Server Logs" *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16 (2016): Web*

[4] H. Li, "A Short Introduction to Learning to Rank" *IEICE Transactions on Information and Systems IEICE Trans. Inf. & Syst. E94-D.10 (2011): 1854-862*

[5] J. Nielsen, "How Users Read on the Web" *Nielsen Norman Group. N.p., n.d. Web. 6 May 2008*

[6] "Lead Section" *Wikipedia. Wikimedia Foundation.* https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Lead_section