PREPRINT
M. Schwarzer, M. Schubotz, N. Meuschke, C. Breitinger, V. Markl, and B. Gipp, "Evaluating Link-based Recommendations for Wikipedia" in Proceedings of the 16th ACM/IEEE Joint Conference on Digital Libraries (JCDL`16), 2016. Downloaded from: www.gipp.com/pub/

Evaluating Link-based Recommendations for Wikipedia

Malte Schwarzer
TU Berlin
ms@mieo.de

Moritz Schubotz
TU Berlin
schubotz@tu-berlin.de

Norman Meuschke
University of Konstanz
norman.meuschke@uni-konstanz.de

Corinna Breitinger
University of Konstanz
isg@uni-konstanz.de

Volker Markl
TU Berlin
volker.markl@tu-berlin.de

Bela Gipp
University of Konstanz
bela.gipp@uni-konstanz.de

ABSTRACT
Literature recommender systems support users in filtering the vast and increasing number of documents in digital libraries and on the Web. For academic literature, research has proven the ability of citation-based document similarity measures, such as Co-Citation (CoCit), or Co-Citation Proximity Analysis (CPA) to improve recommendation quality.

In this paper, we report on the first large-scale investigation of the performance of the CPA approach in generating literature recommendations for Wikipedia, which is fundamentally different from the academic literature domain. We analyze links instead of citations to generate article recommendations. We evaluate CPA, CoCit, and the Apache Lucene MoreLikeThis (MLT) function, which represents a traditional text-based similarity measure. We use two datasets of 779,716 and 2.57 million Wikipedia articles, the Big Data processing framework Apache Flink, and a ten-node computing cluster. To enable our large-scale evaluation, we derive two quasi-gold standards from the links in Wikipedia's "See also" sections and a comprehensive Wikipedia clickstream dataset.

Our results show that the citation-based measures CPA and CoCit have complementary strengths compared to the text-based MLT measure. While MLT performs well in identifying narrowly similar articles that share similar words and structure, the citation-based measures are better able to identify topically related information, such as information on the city of a certain university or other technical universities in the region. The CPA approach, which consistently outperformed CoCit, is better suited for identifying a broader spectrum of related articles, as well as popular articles that typically exhibit a higher quality. Additional benefits of the CPA approach are its lower runtime requirements and its language-independence that allows for a cross-language retrieval of articles. We present a manual analysis of exemplary articles to demonstrate and discuss our findings.

The raw data and source code of our study, together with a manual on how to use them, are openly available at: https://github.com/wikimedia/citolytics

JCDL '16, June 19 - 23, 2016, Newark, NJ, USA

ACM 978-1-4503-4229-2/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2910896.2910908

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, relevance feedback*

General Terms
Information Systems, Recommender Systems, Wikipedia

Keywords
Co-Citation, Co-Citation Proximity Analysis, digital libraries, large-scale evaluations, citation analysis, document similarity measures, link-based, literature recommendations, big data

INTRODUCTION
Literature recommender systems (LRS) are a crucial filtering and discovery tool to manage the vast and continuously increasing volume of documents available in digital libraries and on the Web. Most LRS (approximately 55%) employ content-based document features and corresponding similarity measures to provide recommendations [1].

Especially in academia, LRS are a central fixture among research support tools. Keeping track of the latest research in one's field by identifying the most relevant papers is essential for research progress. The exponentially increasing number of published articles (approximately 1.9 million in 2015[1]) and the increased speed of article availability, e.g. due to Open Access and preprint publishing options, makes thorough literature research even more important, but at the same time more tedious and time consuming for researchers. In academic LRS, citation-based features and document similarity measures have proven valuable [24, 36].

Wikipedia is a large and rapidly growing digital library. As of April 2016, all language-specific versions of the Wikipedia combined contain approximately 39 million articles, of which five million are in English[2]. The English Wikipedia grew by approximately 1,000 articles per day in 2015. All Wikimedia projects received on average 18 billion page views (crawlers excluded) per month in 2015[3]. Despite Wikipedia's size, popularity and rapid growth, little research has addressed the issue of improving information search in Wikipedia through automated generation of article recommendations. Wikipedia relies entirely on manually created and curated links to related articles.

In this paper, we investigate the performance of citation-based similarity measures in recommending related articles in Wikipedia. Our study focusses on comparing the well-established

[1] We estimate the number of articles using a regression model that Bornmann et al. [4] derived from Web of Science data.

[2] http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[3] http://reportcard.wmflabs.org

citation-based similarity measure Co-Citation (CoCit) to its proximity-weighted enhancement Co-Citation Proximity Analysis (CPA). We use Wikipedia links instead of citations to compute the two measures. By including the MoreLikeThis (MLT) function of the Apache Lucene framework, we evaluate a traditional text-based similarity measure employing a term vector space model. MLT was also used in comparable studies [25, 33].

1. BACKGROUND

1.1 Citation-based Similarity Measures

The link-based concepts Co-Citation [31] and Co-Citation Proximity Analysis [11, 12] originate from the field of Library Science. Academic *citations* can be regarded as the offline equivalent of *links* in Wikipedia or on the Web in general. The Co-Citation measure independently proposed by Small and Marshakova-Shaikevic [23, 31] reflects the frequency with which two documents are cited together in other documents. The more frequently two documents are co-cited, the more strongly related they are according to the CoCit measure. Figure 1 illustrates the CoCit concept, where Doc A and Doc B have a co-citation strength of two, since they are co-cited by Doc C and Doc D.

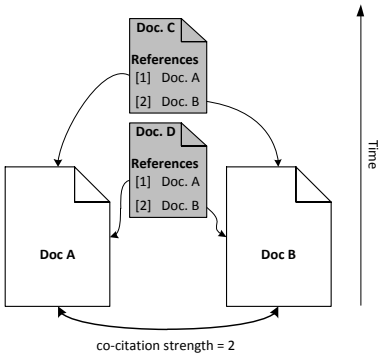


Figure 1: CoCit relationship between documents. Source [8]

The CoCit measure is representative of an era, in which the large majority of academic full texts was not readily available in digital form. Therefore, CoCit exclusively considers entries in the bibliography of academic documents, since this information was accessible using traditional citation indexes [7]. CoCit assigns equal weight to each pair of co-cited documents regardless of *where* in the citing document the citations occur.

Coinciding with the increase in digital availability of academic full texts, Gipp and Beel [12] proposed that taking into account the proximity of co-citations can enhance the effectiveness of the CoCit measure. When the citation markers of co-cited documents appear in close proximity within the citing document, the co-cited documents are more likely to be related. Gipp and Beel coined the concept Co-Citation Proximity Analysis (CPA).

Figure 2 illustrates the CPA approach. The documents B and C are considered more strongly related than A and B, because B is co-cited with C in the same sentence, whereas the citation of document A occurs in a different section of the document. To quantify the degree of relatedness of co-cited documents, CPA assigns a numeric value, the Co-Citation Proximity Index (CPI), to each pair of documents co-cited in one or more citing documents. The CPI reflects the smallest distance between the citation markers of two co-cited documents within a citing document. Gipp and Beel distinguished five levels of co-citation proximity, each of which is assigned a static CPI: same sentence (CPI=1), same paragraph (CPI=1/2), same chapter (CPI=1/4), same journal issue or book (CPI=1/8), same journal, but different issue (CPI=1/16). The CPA score is formed by summing up the proximity weighted co-citations over all co-citing documents.

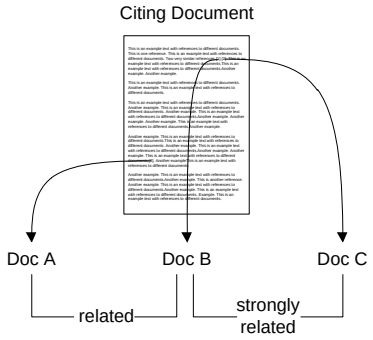


Figure 2: Document similarity assessment according to Co-Citation Proximity Analysis. Source [8]

1.2 Text-based Similarity Measures

To compare the results of the link-based similarity measure, we use the text-based MoreLikeThis (MLT) function of Apache Lucene. MLT uses a Vector Space Model (VSM) of terms as introduced by Salton, Wong and Yang [30] in combination with Term Frequency-Inverse Document Frequency (TF-IDF) weighting proposed by Jones [19]. Several studies and its widespread use among websites have proven MLT's suitability for determining website similarity [5, 20, 29].

2. RELATED WORK

Several publications investigate the placements of citations within the full-texts of documents as additional information for co-citation analysis. Tran et al. [33] and Eto [6] showed the increase in effectiveness of employing sentence-level, over paper-level, citation proximity on the task of retrieving related articles. Liu and Chen [22] analyzed the distribution of co-citations at four levels of proximity: article, section, paragraph and sentence level. They found that sentence-level co-citations can increase the accuracy and efficiency of co-citation analysis.

Gipp et al. investigated the analysis of citation patterns in academic documents to identify disguised forms of academic plagiarism such as paraphrases or translations. They proposed several detection algorithms, which aside from citation proximity also consider the order of citations, citation counts, and other properties to identify suspicious citation patterns [13]. They demonstrated the effectiveness and efficiency of their approach "Citation-based Plagiarism Detection" by analyzing known plagiarism cases [10] and by discovering previously unknown cases in a large full text collection [9].

Recommending citations for academic papers is a well-researched problem. Early approaches used collaborative filtering [24] or (co-)citation information [32]; hence, they required author profile information or partial bibliographies. More recent works focus on citation context analysis to identify suitable citations for specific parts of a paper. For instance, He et al. [14] trained a probabilistic topic model for the citation context, i.e. text range, surrounding user-specified placeholders for citations to identify and rank documents most relevant to the topic of the citation context. Huang et al. [17] extended this approach by proposing a translation model to map the topic of citation contexts to the topics of potential sources while accounting for differences in the vocabulary of the citation contexts and the sources. Most recently, Huang et al. [16] used a probabilistic neural network to model the relationship between citation contexts and potential sources.

For finding related pages in the English Wikipedia, Ollivier and Senellart [25] proposed the Green Measure, which uses Markov chains, and compared the measure to other methods. They found that Green Measure has both the best average results and the best robustness compared to Co-Citation, Cosine similarity with TF-IDF weighting and PageRank [26] of links. Aside from this work, we are unware of research addressing the recommendation of related pages in Wikipedia.

So far, the performance of the CPA approach has been evaluated for academic citations, but not for a large-scale hyperlinked environment, such as the English Wikipedia corpus. Additionally, no large-scale evaluation using "See also" links and clickstreams as quasi-gold standards has yet been performed on Wikipedia.

3. EXPERIMENTAL SETUP

3.1 Test Collection

Our test collection is a dump of the English version of Wikipedia. The dump was created in September 2014, consists of 4.6 million Wikipedia articles in XML Wiki markup, and has a size of 99 GB. To get an overview of the test collection's composition and to enable a comparison with other collections, we collected information on article length and the number of in-links. Figure 3 shows the distribution of words and in-links among articles. The word frequencies are grouped into bins of 20 words.

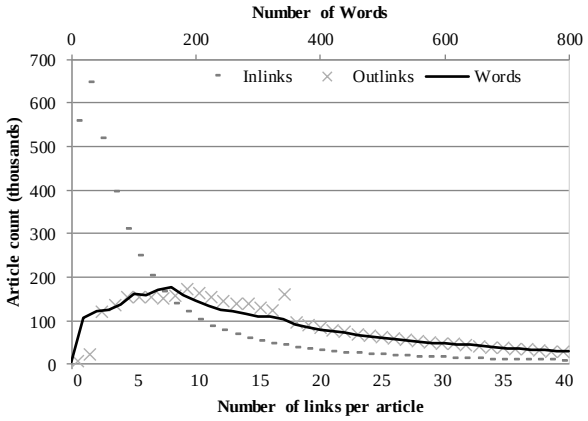


Figure 3: Distribution of word frequency (solid), in-links (dashes), and out-links (crosses) of articles.

On average, articles contained 740.54 words. The longest article contained 75,178 words. There is a consistently strong correlation between the number of out-links and the number of words for all article lengths. The distribution of in-links is heavily skewed. About 1.7 million of the 4.6 million articles have less than three in-links. On average, an article had 20.5 in-links. The most linked to article was "United States", which received 392,494 in-links. As reported by Belomi and Bonato [3], Wikipedia articles with a high number of in-links are mainly about geopolitical topics, famous people, abstract nouns, or common words.

3.2 Information Need

Our goal was a large-scale evaluation of the performance of similarity measures in recommending related Wikipedia articles. Instead of selecting a number of topics and defining topic-specific information needs, we wanted to obtain an understanding of how well the methods perform for the entire Wikipedia with its vast range of topics. Therefore, we defined a generalized information need for our study as follows:

"Recommend related Wikipedia articles that may be of interest to a reader of the source article".

3.3 Quasi-Gold Standards

Given the large scope of our study, we required human relevance judgments that suit our information need, are available for large parts of the collection and a broad range of topics, and are obtainable in an automated fashion. We derived two quasi-gold standards satisfying these requirements from analyzing (a) "See also" links and (b) clickstream data. In contrast to a traditional user study, which is typically limited to a few hundred articles at most, these datasets allowed an evaluation for 779,716 articles using "See also" links and 2.57 million articles using the clickstream data set.

Nonetheless, this evaluation approach lacks completeness. "See also" links and clickstream data are only approximations of complete relevance judgments. Therefore, we refer to them as quasi-gold standards, not gold standards. A quasi-gold standard is an approximation of a 'perfect' reference model. Both quasi-gold standards are being applied for the first time, and have yet to be evaluated by the research community.

3.3.1 "See Also" Links

A unique characteristic of Wikipedia articles is not only that they contain links to additional information in the form of internal references or external links, but also that they contain so-called "See also" sections. The purpose of these sections is to provide links to topically related Wikipedia articles [35], which results in these links acting as recommendation sets for relevant literature. Correspondingly, "See also" links are equivalent to a quasi-gold standard that allows a performance evaluation of a recommendation system.

Therefore, we classified articles as relevant if the retrieved article is listed in the "See also" section and as irrelevant otherwise. However, it is in this second assumption that we see a problem: We expect the "See also" links to be an incomplete quasi-gold standard created by a few Wikipedia editors. We assume that the main objective of Wikipedia editors lies in creating textual content, rather than providing useful literature recommendations, which means that if a retrieved document is not included in the "See also" links, it can still be topically related, i.e. relevant. Therefore, we can only decide if a result is relevant, but not if it is irrelevant. A true binary classification is not possible.

Hence, we expect a precise true positive classification for articles that exist as "See also" links. However, many results could be classified as false negatives, even if a result is truly relevant because the recommendation is missing in the "See also" links.

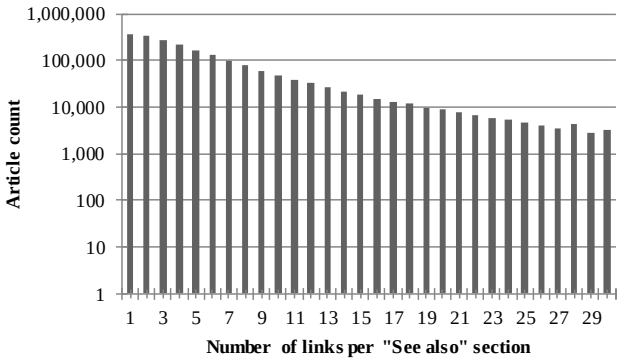


Figure 4: Number of links per "See also" section. Avg.: 2.6 links. Total: 2,028,146 links.

We extracted the "See also" section and its links using an automated process. Figure 4 shows the distribution of the number of "See also" links in the Wikipedia dataset. The test collection contained 779,716 Wikipedia articles with "See also" sections (17% of the corpus), where each section on average contained 2.6 links. This low number of links per section additionally contributed to the incompleteness of relevance judgments, since the number of relevant recommendations that could be made from the Wikipedia corpus is likely greater than the number of available "See also" links.

3.3.2 Clickstreams

The recent publication of Wikipedia clickstreams by WikiResearch [28, 37] allowed us to use a second quasi-gold standard. The dataset contains clickstreams for 2,572,063 articles (56% of the corpus) in the form of aggregated HTTP referrer information during the month of February 2015. The HTTP referrer indicates the page from which a user clicked to the article in question. Using this data, we can determine the number of clicks on out-links for articles. For out-links, which occur multiple times in an article, only the total number of clicks is provided. WikiResearch cleaned the dataset from computer-generated clicks (bot activity). However, researchers of the Wikipedia foundation have observed that the filtering of bots should be improved [34]. We assume that the dataset contains some noise from bot activity, but we cannot quantify or reduce the noise level, since only aggregated clickstream data was available to us. In the future, WikiResearch plans to release more datasets, which would increase the value of clickstreams as a quasi-gold standard.

We consider the number of clicks on a link as a cardinal relevance classification regarding the linked article. The more often a link is clicked, the more relevant we assume the article to be. Whether this assumption holds true for *all* articles, and whether it is the major force driving clicks, has not been proven. Other factors can also affect the number of clicks, such as the descriptiveness value of the link, or the link's position within the article. A recently published study showed that the Click-Through-Rate decreases in proportion to the link's position from the top [27].

The two quasi-gold standards differ in their conceptual properties: While the "See also" quasi-gold standard is created by the Wikipedia editors; clicks are relevance judgments by all readers. Moreover, clicks can only occur on links that exist in the article content. Such in-content links are also included for navigational purposes, while "See also" links are exclusively literature recommendations. The Wikipedia manual states to only add links in "See also" sections that do not exist in other parts of the article.

3.4 Performance Measures

Each quasi gold standard is evaluated separately to ensure that all Wikipedia articles contribute equally to the results, independent of an article's number of "See also" links or its popularity.

In the "See also" evaluation, we use the rank-based Mean Average Precision (MAP) score (equation 1) to quantify recommendation quality. MAP represents the mean of the average precision scores for a set of queries Q (In our case: Wikipedia articles). $R_{q,j}$ denotes a relevant result for query q retrieved at rank j. We calculate MAP for the 10 top-ranked results, i.e. k=10. All articles are weighted equally in the final MAP score regardless of the article's number of "See also" links.

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{|R_q|}\sum_{j=1}^{|R_q|}\mathrm{Precision}(R_{q,j}) \quad (1)$$

We also performed test runs that calculated the performance measure Mean Reciprocal Rank (MRR) in addition to MAP during the "See also" evaluation. MRR represents the rank position of the first relevant result averaged over all queries. Evaluating the approaches according to MAP or MRR yielded no significant differences in the performance relation of the approaches. Therefore, we chose to only report MAP results in this paper, since we consider MAP as more representative of the performance of an approach with regard to all results. The code to calculate MRR is included in the GitHub repository for this paper.

In the clickstream evaluation, we measure recommendation performance using the Click-Through-Rate measure (CTR) (equation 2) for the top-k-results with k set to 1, 5, and 10 respectively. CTR represents the ratio of clicks $C_{s,d}$ on a link from article s to article d and the number of all outgoing clicks for article s [18]. Popular Wikipedia articles can generate more clicks than niche articles. Nevertheless, we followed the approach of equally valuating each article independent of its popularity.

$$\mathrm{CTR}(s,d) = \frac{C_{s,d}}{\sum_{j=1}^{|C_s|}\mathrm{C_{s,j}}} \quad (2)$$

3.5 Implementation

For the sake of transparency and to improve reproducibility in recommender system research [2], we have published the data and source code used in our study together with a manual on GitHub: https://github.com/wikimedia/citolytics

3.5.1 More Like This

To generate the MoreLikeThis result set, we used a Java application and an Elasticsearch cluster. The application consists of four sub-tasks: extracting all articles from the Wikipedia XML dump, adding them to the Elasticsearch index, performing MoreLikeThis queries and storing all results as CSV.

3.5.2 CPA

We implemented the CPA algorithm as an Apache Flink job [21] in Java. In contrast to MLT, CPA does not require an indexing process. Instead, the CPA results are directly generated from the Wikipedia XML dump. This requires extraction of the full link graph and performing CPI computation. These operations are expressed in the MapReduce programming model. For completeness, we also resolve redirections for Wikipedia links that do not point directly to their destination.

The static classification of CPI values originally proposed by Gipp and Beel [12] (see Section 1.1) does not fit our test collection. Wikipedia articles are not organized in journals, nor do they follow the structure of scientific documents. Thus, we introduce a new dynamic model of CPI that can be adjusted depending on the requirements of the test collection.

We considered the proposal of Tran et al. [33] to generalize the citation proximity level. Analogous to the Term-Document Matrix, used in text-based approaches like VSM, we define the Link-Position Matrix $v_{i,j}$ of dimension $m \times m$ that stores the link position for all m documents. Specifically, the column for document j, $v_{*,j}$ holds the positions for links to other documents in words counted from the beginning of the document. Without loss of generality, we assume that a document links only once to another document. This complies with the conventions for authoring Wikipedia articles, which state that only the first mention of a concept should be linked.

Thus, for our use case, we redefined CPI as:

$$\mathrm{CPI}(a,b) \equiv \sum_{j=1}^{m} \Delta_j(a,b)^{-\alpha}, \quad (3)$$

$$\text{with } \Delta_j(a,b)^{-\alpha} = \begin{cases} \left|v_{a,j} - v_{b,j}\right|^{-\alpha} & v_{a,j} > 0 \wedge v_{b,j} > 0 \\ 0 & \text{otherwise} \end{cases}.$$

This definition states that for a document pair (a, b), the CPI is the sum of the proximity of their co-citations Δ_j, where the proximity is the link-distance damped by an exponential tuning parameter α, which determines the influence of the distance. The value of α needs to be computed depending on the document type, i.e. the model needs to be optimized. Note that negative values for α are counter-intuitive, because a negative value of α would result in a weighting that prefers co-citations with a greater distance. Furthermore, the case of $\alpha = 0$ implies:

$$\mathrm{CPI}(a,b) = \sum_{j=1}^{m} \begin{cases} 1 & v_{a,j} > 0 \wedge v_{b,j} > 0 \\ 0 & \text{otherwise} \end{cases}. \quad (4)$$

In this specific case, CPI is independent from link distance and equivalent to CoCit, since only the number of co-citations is counted, i.e. proximity has no effect.

3.5.3 "See Also" Evaluation

We collected the data for the "See also" quasi-gold standard from the Wikipedia dump by looking up sections titled "See also" and extracting the sections' links. We merged the resulting dataset with the MLT and CPA results using the article name. Lastly, we ensured that a "See also" link existed for each retrieved article.

3.5.4 Clickstream Evaluation

The data required for the clickstream evaluation was obtained from Wikiresearch as a CSV file. Therefore, no pre-processing was required. We assigned the clickstream data to CPA and MLT results, i.e. we assigned each article recommendation the respective number of clicks on the link and its CTR. In the final evaluation process, we merged all result sets with the corresponding quasi-gold standards.

3.5.5 Computing Infrastructure and Runtime

The experiment was performed on a cluster of 10 IBM Power 730 (8231-E2B) servers. Each machine had two 3.7 GHz POWER7 processors with 6 cores (12 cores in total), 2 x 73.4 GB 15K RPM SAS SFF Disk Drive, 4 x 600 GB 10K RPM SAS SFF Disk Drive and 64 GB of RAM.

Table 1: Approximated runtimes for each task.

Task	Runtime
MoreLikeThis (Elasticsearch)	
Indexing	7h 30min
Retrieval	53h 45min
CPA (Apache Flink)	
Computing Results	7h 45min
Evaluation (Apache Flink)	
"See also"-links	45min
Clickstream	50min

We used Apache Flink v0.8 (2015-01-19) and Hadoop v2.4.1 (2014-06-21). The text-based similarity measure was evaluated using Elasticsearch v1.4.2 (2014-12-16). All versions were the latest stable releases at the time of the experiment. We used the software's default settings, i.e. neither Apache Flink nor Elasticsearch had been optimized for runtime performance. Although we did not focus on runtime performance and none of the tested document similarity measures had been optimized, the difference in runtime between CPA and MLT, as listed in Table 1 shows that MLT involves a more extensive computation than CPA. This is conceptually obvious, since the data volume for the recommendations based on words vs. links differs significantly. Also, MLT requires additional cleaning techniques such as stop word removal and TF-IDF weighting.

4. RESULTS

4.1 Optimizing the CPI Model

Since we use a dynamic CPI model instead of the static CPI values used in the original approach by Gipp and Beel (Section 1.1), we need to adjust CPA for Wikipedia articles before benchmarking the approach. We need to find a value for the constant α that achieves the best MAP score for the "See also" evaluation and the best CTR score for the clickstream evaluation. Since our goal is to optimize α specifically for the Wikipedia collection, we use the full collection instead of splitting up the collection into a training and test dataset. The later procedure would be appropriate if we were searching for an α value that performs best for different collections.

To find the value for α that performs best for our collection, we applied CPA with α values from -1 to 5 in 0.01 increments. Then, we evaluated the retrieved top-k results with k=10 of each batch by calculating the MAP and CTR scores (Figure 5). CPA performed best in terms of MAP with α set to 0.81 and in terms of CTR with α set to 0.90 (see marks in Figure 5). Thus we used these optimized α values in the corresponding CPI models during the "See also" and clickstream evaluation.

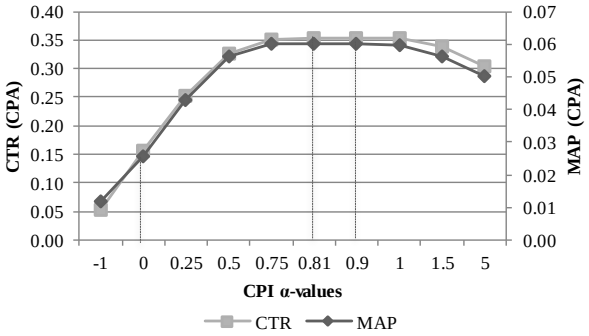


Figure 5: CTR and MAP scores for CPA with various CPI α-values. MAP_{max} at α = 0.81, CTR_{max} at α = 0.9

Moreover, the graph in Figure 5 proves the consistently lower performance of CoCit compared to CPA. CoCit is a special case of CPA with α set to zero (left mark in the graph). Only for negative α values, CoCit performs better than CPA. Using negative α values would cause CPA to assign higher scores to more distant co-citations, thereby effectively reversing the concept of the CPA measure and reducing CPA's performance. The graph therefore proves the benefit of assigning higher scores to co-citations at closer proximity.

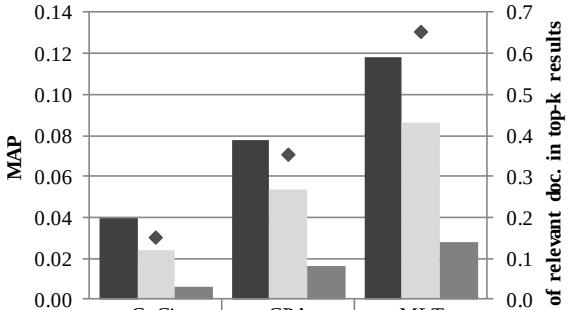
4.2 Evaluation for Quasi Gold Standards

In the following, we present the evaluation results for the two quasi-gold standards presented in Section 3.3. To be included in the "See also" evaluation, a Wikipedia article must contain a "See also" section, which was true for 779,716 articles. To be included in the clickstream evaluation, clickstream data had to be available for the article in question, which was true for 2,572,063 articles.

To enable optimal comparability of the evaluated similarity measures, the following sections report results for a "unified dataset", i.e. those articles, for which all three evaluated measures retrieved the same number of related articles. For example, CoCit and CPA cannot generate recommendations for articles without in-links, hence we excluded such articles from the unified dataset. This procedure reduced the dataset for the "See also" evaluation from 779,716 articles to 659,642 articles (-120,074) and the dataset for the clickstream evaluation from 2,572,063 articles to 2,535,987 articles (-36,076). To ensure that unifying the datasets did not skew the evaluation, we calculated all performance scores for CoCit, CPA and MLT also based on the sets of all related articles that the measures could identify. The maximum difference in any score was 1.3% (average number of clicks for CPA) and for most scores less than 1% compared to the results of the unified dataset. The GitHub repository for this paper includes the results for the unified dataset and the results for set of all related articles.

4.2.1 "See Also" Links

Figure 6 shows that MLT performed better than CPA in terms of MAP and the average number of retrieved relevant documents, while CoCit performed worst. The MAP score of CPA is less than half of MLT's score; CoCit's score is less than a quarter of MLT's score. The average number of relevant documents of MLT and CPA tripled from k=1 to k=5 and nearly quadrupled from k=1 to k=10. We expected significant performance differences between CoCit and CPA, since the CPI optimization already showed that CoCit is an under-performing variation of CPA. On the other hand, we see an advantage of text-based MLT over the citation-based similarity measures, when judging recommendation relevance using "See also" links.



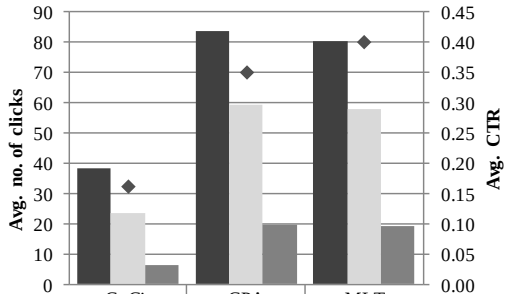
	CoCit	CPA	MLT
k=10	0.20	0.39	0.59
k=5	0.12	0.27	0.43
k=1	0.03	0.08	0.14
MAP	0.03	0.07	0.13

Figure 6: Results of "See also" link-based evaluation.

4.2.2 Clickstreams

Figure 7 shows the CTR ranking of the clickstream evaluation. CPA accounted for more clicks than MLT for any value of k. MLT achieved the highest CTR, however, the ratio of the CTR scores of MLT and CPA (1.13) was significantly lower than the ratio of the MAP scores of the two approaches (1.92). CoCit again performed worst with regard to both scores.

The improved performance of CPA in this evaluation compared to the "See also" evaluation indicates that CPA performs better than MLT for popular articles, while MLT is more effective for niche articles. In the following, we present possible interpretations for this observation, which, however, need further investigation.



	CoCit	CPA	MLT
Clicks (k=10)	38.34	83.87	80.64
Clicks (k=5)	23.52	59.50	58.00
Clicks (k=1)	6.39	19.61	19.08
CTR (k=10)	0.16	0.35	0.40

Figure 7: Results of clickstream evaluation. MLT yields best CTR, while CPA generates the most clicks.

Popular articles typically attract many visitors and thus have a larger impact on the total click count than niche articles. However, CTR values every article equally, thus CTR does not reflect the comparably better performance of CPA for popular articles as strongly as the average number of clicks.

Popular articles also tend to have more co-authors. Therefore, the collaboratively generated 'link set' contained within popular articles might be of higher relevance, thus generating higher numbers of clicks and CTRs. To be able to support this hypothesis, we would need to evaluate the performance with regard to indicators of article quality [15].

Additionally, popular articles likely receive more in-links, which affects CPA's performance. We further investigate this property in Section 4.3.2. Another cause for CPA performing better for popular articles might be that bots, i.e. computer generated clicks, have a proportionally larger impact on niche articles. Consequently, the quality of the quasi-gold standard for these articles might be lower than for articles of average popularity. As we explain in Section 3.3.2, we cannot quantify this effect, since the data we used had been aggregated, thus preventing us from filtering bots on our own.

4.3 Article Properties

In this subsection, we provide details on the evaluation of CPA and MLT depending on article properties, such as the number of words and in-links. We omit CoCit in this evaluation, since the previous "See also" and clickstream evaluations already showed its inferior performance compared to CPA.

Figure 8 and Figure 9 show the performance in terms of MAP and CTR with respect to words and in-links. The graphs do not cover the full corpora: For the sake of visibility we do not plot results for articles with more than 3,000 words (9.07% of the articles in the in "See also" dataset, 5.90% of the articles in the clickstream dataset) or 400 in-links (2.66% of the articles in the "See also" dataset, 1.22% of the articles in the clickstream dataset).

4.3.1 Words

The performance plot with respect to article length, see Figure 8, reveals some interesting results. First, we see that MLT consistently performs better than CPA, when using MAP, but when using CTR, the performance ranking varies depending on the number of words. For articles with less than around 1,400 words MLT is superior, otherwise CPA performs slightly better.

Second, MLT's and CPA's MAP and CTR graphs show similar tendencies, but with one exception: MLT's MAP and CTR scores for very short articles (30-50 words) are exceptionally high, but drop sharply for slightly longer articles (60-150 words). For articles with more than approximately 150 words, MLT's MAP and CTR scores increase steadily and peak at article lengths of approximately 250 words. For articles longer than 250 words, MLT's MAP and CTR scores steadily decline. CPA's MAP and CTR scores, on the other hand, increase with article length up to lengths of approximately 400 words. Beyond this point, the CPA's MAP and CTR score remain relatively stable.

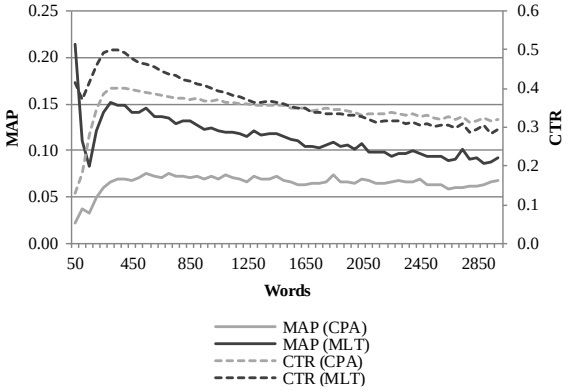


Figure 8: Performance evaluation in relation to number of words per article.

MLT's performance is more strongly affected by article length than CPA's. Short articles simply offer less data for both a text-based and link-based similarity assessment. If an article contains few words, it is difficult to determine topic-defining keywords and find other articles with matching topics. Short articles also typically have fewer in-links, e.g., because they are stubs. Therefore, both MLT and CPA require an article length of approximately 250 or more words to perform well. MLT's MAP peak for articles with around 50 words is an outlier phenomenon. Such very short articles normally contain only a single sentence on one topic, a list, a table, or specific vocabulary. Therefore, such articles often allow an accurate text-based similarity assessment.

While CPA reaches a relatively stable performance in terms of MAP and CTR, MLT's MAP and CTR score decline steadily for articles with 450 words or more. Long articles often cover several subtopics, which decrease the performance of VSM-based text similarity approaches like MLT. The vocabulary of subtopics can vary, thus making it difficult to determine a set of words that represents the breadth of topics present in the article. CPA's performance is hardly affected by article length, given a critical mass of in-links has been reached. This result is intuitive given that CPA's performance exclusively depends on in-links.

4.3.2 In-Links

Figure 9 shows the plot of MAP and CTR scores depending on the number of in-links. Both MLT and CPA performed best for approximately 20 in-links. For more in-links, the performance declines steadily as the number of in-links increases. This plot also shows a change in the CTR performance ranking of CPA and MLT. For less than 50 in-links MLT performs better; for more than 50 in-links CPA performs better. On the contrary, the ranking according to MAP does not change.

In-links as a data source are essential for link-based similarity measures, but do not directly affect text-based similarity measures. Seeing MLT perform better than CPA in terms of CTR for articles with less than 20 in-links is therefore intuitive. It is also intuitive that CPA's CTR scores increase as the number of in-links increases in the range of 0 to 20 in-links.

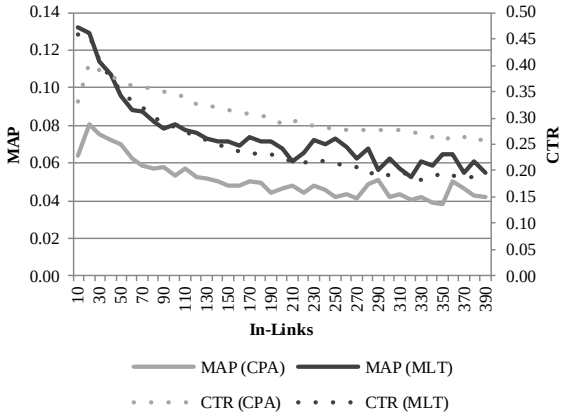


Figure 9: Performance evaluation considering number of in-links per article.

The reason that CPA's CTR scores peak at 20 in-links and decline thereafter and MLT's CTR scores decline steadily as the number of in-links increase may not be as intuitive. We attribute this behavior to the nature of articles that receive many in-links. Such articles typically cover broad topics, e.g. countries, as Belomi and Bonato [3] also reported before. We explain in Section 4.3.1 that text-based similarity measures like MLT perform comparably worse for such articles than for articles with narrowly similar topics. Figure 9 demonstrates that also link-based measures like CPA perform worse for broad-topic articles, because such articles receive in-links from many and topically diverse articles. This diversity of received in-links reduces the likelihood that the article in questions is frequently co-cited in closer proximity with other articles, hence reducing the performance of CPA.

4.4 Manual Sample Examination

To test the validity of "See also" links and clickstreams as gold standards, we manually evaluated a small and random subset of corpus articles. From these articles, we present and discuss the three exemplary articles shown in Tables 2 - 4. We chose the articles for their diversity and comprehensibility. Tables 2 - 4 show the recommendations of CoCit, CPA and MLT with the corresponding rank, similarity score for each measure in parentheses, and click counts. Recommendations that are part of the "See also" links are underlined.

4.4.1 Technical University of Berlin

The article about the Technical University of Berlin (TUB) includes information about the university's history, campus, organization, and a list of notable alumni and professors.

Both link-based measures retrieved two documents, which were included in the "See also" links and received clicks (HU-Berlin and FU-Berlin, underlined in Table 2). The MLT results have a clear focus on "University" as the topic, since all recommended articles are about universities, but from other cities and countries.

In this case, it can be said that the best results were produced by the CPA algorithm, followed by CoCit and MLT. While the CPA results can all be considered relevant, the MLT approach in particular produced a list of irrelevant institutions. For example, the University of Economics Varna in Bulgaria, or the Technological University Hpa-An in Myanmar.

Table 2: Results for "Technical University of Berlin"

	CoCit result	Clicks	CPA result	Clicks	MLT result	Clicks
1	Germany (660)	0	Germany (20.0)	0	Technical University of Sofia (0.86)	0
2	Berlin (487)	20	Berlin (17.6)	20	University of Economics Varna (0.74)	0
3	Humboldt University of Berlin (245)	42	Humboldt University of Berlin (10.0)	42	Vilnius College of Technologies and Design (0.64)	0
4	Ludwig Maximilian Uni. of Munich (229)	0	RWTH Aachen University (8.4)	0	Braunschweig University of Technology (0.63)	0
5	World War II (178)	0	Technische Universität München (5.9)	0	Technical University of Gabrovo (0.60)	0
6	United States (174)	0	Charlottenburg (5.6)	0	Chemnitz University of Technology (0.59)	0
7	RWTH Aachen University (172)	0	Mathematics (5.3)	0	Technische Universität Ilmenau (0.56)	0
8	Free University of Berlin (170)	0	Free University of Berlin (4.9)	0	Technical University of Dortmund (0.51)	0
9	Heidelberg University (142)	0	Habilitation (4.3)	0	Dresden University of Technology (0.50)	0
10	Mathematics (139)	0	Ludwig Maximilian Uni. of Munich (3.8)	0	Technological University Hpa-An (0.49)	0

"See also" links: Hertie School of Governance, Berlin University of the Arts, Free University of Berlin, Humboldt University of Berlin, Berlin School of Economics and Law, Beuth University of Applied Sciences Berlin

Total clicks: 596

Table 3: Results for "Elvis Presley"

	CoCit result	Clicks	CPA result	Clicks	MLT result	Clicks
1	AllMusic (5977)	0	The Beatles (115.5)	247	Sun Studio (1.16)	0
2	The Beatles (5030)	247	Frank Sinatra (61.9)	139	From Elvis in Memphis (1.14)	516
3	Billboard magazine (4425)	0	Johnny Cash (53.2)	140	List of songs recorded by Elvis Presley on the Sun label (1.10)	240
4	United States (3146)	52	Jerry Lee Lewis (50.7)	73	Peter Guralnick (0.99)	0
5	Frank Sinatra (2756)	139	RCA Records (45.1)	175	Colonel Tom Parker (0.96)	1175
6	The Rolling Stones (2374)	0	Rock and roll (42.9)	306	The Blue Moon Boys (0.94)	100
7	Billboard Hot 100 (2203)	12	Heartbreak Hotel (38.0)	720	Elvis Presley's Army career (0.89)	619
8	Johnny Cash (2157)	140	Jailhouse Rock song (36.4)	260	Jailhouse Rock film (0.87)	1132
9	Cliff Richard (1996)	0	Roy Orbison (34.8)	96	I Want You, I Need You, I Love You (0.86)	83
10	Bob Dylan (1930)	77	United States (30.4)	52	Elvis Presley albums discography (0.83)	6084

"See also" links: Honorific nicknames in popular music, Elvis Presley Enterprises, List of best-selling music artists, Personal relationships of Elvis Presley, List of artists by number of UK Albums Chart number ones, List of artists by total number of UK number one singles

Total clicks: 92,379

Table 4: Results for "Newspaper"

	CoCit result	Clicks	CPA result	Clicks	MLT result	Clicks
1	United States (4130)	0	Broadsheet (428.0)	59	The Daily Courier Arizona (0.90)	0
2	Broadsheet (2569)	59	Magazine (331.5)	119	Online newspaper (0.88)	142
3	English language (1732)	0	Tabloid newspaper format (246.7)	35	History of British newspapers (0.86)	168
4	Tabloid newspaper format (1690)	35	United States (225.4)	0	List of newspapers in the United States by circulation (0.86)	0
5	Race and ethnicity in the United States Census (1257)	0	Publishing (102.2)	0	Newspaper circulation (0.84)	23
6	The New York Times (1041)	118	English language (96.2)	0	Midland Daily News (0.78)	0
7	New York City (890)	0	Journalist (86.2)	32	The Huntsville Times (0.77)	0
8	World War II (831)	0	Book (80.3)	11	Decline of newspapers (0.75)	0
9	Magazine (822)	119	Comic strip (80.0)	37	The Leaf-Chronicle (0.74)	0
10	United Kingdom (805)	0	Radio (78.9)	0	The Ann Arbor News (0.74)	0

"See also" links: List of newspaper comic strips, Lists of newspapers

Total clicks: 4,516

The universities considered relevant by the CPA approach are all well-known Universities in the region with a strong technical focus, similar to the technical university of Berlin.

The poor performance of MLT in this case can be explained by the weakness of text-based approaches where a strong emphasis lies on similar words in the documents. Text describing a university is usually similar, given that generic characteristics such as the number of students, etc. is described, which automatically leads to a "high" similarity. Possibly, Wikipedia authors reused text when writing the article about the university in Burma. Citation-based approaches are not affected by text reuse.

4.4.2 Elvis Presley

The biographical Wikipedia article about the American singer and actor Elvis Presley is relatively long. The article contains 24,298

words, received 5,834 in-links and provided 92,379 out-clicks.

None of the articles recommended by any approach were part of the "See also" links, but most recommendations are related to the topic. The topics recommended by CoCit and CPA are broader than the results of MLT. Furthermore, CoCit's recommendations for the articles "AllMusic", an online music database, and "Billboard magazine" are notable: Even though both articles are music-related, they lack a direct connection to Elvis Presley. These recommendations were caused by links that did not belong to the actual article text, e.g. infoboxes or the article footer.

4.4.3 Newspaper

The "Newspaper" article contains general information on newspapers as periodical publications, their historical development, their categories, formats, and other newspaper related topics. The article consists of 6,313 words and is linked by 7,611 other articles. The "See also" section includes two links to newspaper related lists: "List of newspaper comic strips" and "Lists of newspapers".

MLT, CPA, and CoCit all failed to retrieve any of the "See also" links, which is not surprising, since the only two "See also" links linked to another list. Despite all articles retrieved by MLT being newspaper related, they were also overly narrow and irrelevant for the broad and internationally-oriented 'Newspaper' article. MLT recommended articles on actual newspaper publications, e.g. "The

Daily Courier Arizona", or "Midland Daily News"; However, these publications are so provincial, that they will be irrelevant to most readers. CPA, on the other hand, retrieved a broader spectrum of related topics, for example newspaper formats ("Tabloid", "Magazine", "Broadsheet") or other media ("Book", "Comic strip", "Radio"). Two of CPA's results ("United States" and "English language") were not topically relevant. CoCit retrieved many irrelevant articles from the geopolitical category ("United States", "New York City", etc.).

4.4.4 Summary Manual Evaluation

The results presented for these three examples were typical of other articles examined. MLT tended to retrieve topically more narrow articles compared to the citation-based approaches. CPA usually produced more relevant recommendations than CoCit. We observed that the recommendations were of a different nature for each approach. While CPA's recommendations were consistently plausible, MLT had the tendency to recommend obscure articles. For example, MLT recommended a University in Myanmar (Technological University Hpa-An) for the article 'Technical University Berlin' or an internationally virtually unknown newspaper ('The Daily Courier Arizona') at rank 1.

The result of the manual evaluation showed that CPA recommends topically broader articles, but with consistent relevance compared to the often niche results of MLT. However, because this evaluation approach is highly subjective and dependent on a user's specific information need, we invite the reader to examine the examples in the Tables 2 - 4 as well as additional results available in the repository to make a judgement.

5. DISCUSSION

In the "See also" evaluation, the text-based MLT measure retrieved more related articles and achieved higher MAP than both link-based measures. CPA followed at second rank and clearly outperformed the third-ranked CoCit in this evaluation. Links outside of the article text, e.g., in information boxes or article footers, were a source of irrelevant CoCit and CPA results, since such links are commonly less related to the article's topic.

For example, in the article on Elvis Presley, CoCit identified the link to the "AllMusic" category at the top rank. Devaluating or ignoring these links in future studies should improve the performance of the link-based similarity measures. Such a procedure would correspond to the stop word removal in MLT. For the CPA approach, adjusting the CPI weighting scheme could reduce the effect of such Wikipedia-specific unrelated results. For instance, the quantification of citation proximity should be adjusted for article length or the number of in-links an article receives. Such a normalization can devalue links to general articles that are frequently co-cited but often have no topical relevance, e.g. geopolitical articles such as "United States".

Not surprisingly, articles that CPA retrieved as relevant consistently achieved the highest number of clicks in the clickstream evaluation. MLT followed at second rank and CoCit at third rank in this regard. Yet, MLT achieved slightly higher CTR scores than CPA in this evaluation, with CoCit again following at rank three.

These results indicate that traditional text-based methods are a well-performing "general purpose" approach for recommending related Wikipedia articles regardless of specific article properties. CPA is better suited to retrieve popular articles. Due to Wikipedia's collaborative approach to article curation, popular articles are also typically longer and of higher quality.

Manually examining samples also indicated that CPA and MLT have different strengths that are not adequately reflected by the "See also" quasi-gold standard. The link-based approaches, especially CPA, tended to retrieve articles from a broader range of related topics than MLT. For instance, for the query "Newspaper" MLT mostly recommended actual newspapers, e.g. "The Daily Courier Arizona". CPA on the other hand retrieved more generally related topics, e.g., newspaper formats such as "Tabloid" or "Broadsheet". In our perception, CPA and MLT performed similarly well in identifying related articles, yet the type of relatedness differed.

Two advantages of the link-based measures over the text-based measure are their significantly lower runtime requirement (see Table 1) and their language-independence. Citation or link analysis can be performed for texts in any language and can also be employed for retrieving texts across languages. Text-based measures like MLT are language-dependent.

Summarizing our findings, we conclude that the advantageousness of the link-based over the text-based approach depends on the information need of the user. If a user is interested in articles that address a specific topic, in a single language and from a relatively narrow perspective, text-based recommendations likely suit the user's needs better than link-based recommendations.

If the user desires a broader overview of a topic, and also wants to retrieve articles in different languages, or if the user values factors, such as article popularity and quality, then link-based recommendations fulfill these requirements better than text-based recommendations. Ultimately, a combined approach that includes link-based, text-based and potentially other document similarity measures is likely to achieve the best recommendation quality.

6. CONCLUSION

This paper introduced the first implementation of Citation Proximity Analysis (CPA) for a hyperlink environment use case. We adapted the CPA's Citation Proximity Index (CPI) from the academic literature domain, i.e. citation analysis, to the analysis of links. Subsequently, we performed a large-scale evaluation of the performance of the adapted CPA approach, Co-Citation (CoCit), and Apache Lucene's MoreLikeThis (MLT) function for recommending related documents in two datasets of 779,716 and 2.57 million Wikipedia articles. We used the Big Data processing framework Apache Flink and a ten-node computing cluster to compute article similarities for each approach.

To perform this large-scale evaluation, we introduced two novel quasi-gold standards: the links in Wikipedia's "See also" sections and a comprehensive clickstream dataset as estimators of the relevance for Wikipedia articles.

We found that the link-based and text-based approach to recommending articles in Wikipedia have complementary strengths. The text-based MLT method performs well in identifying closely related articles. The CPA approach, which consistently outperformed CoCit, is better suited for identifying a broader spectrum of related articles as well as popular articles that typically exhibit a higher quality. Additional benefits of the CPA approach are its lower runtime requirements and its language-independence, which allows cross-language retrieval of articles.

Our findings suggest that an approach that combines link-based, text-based, and potentially other recommendation algorithms, shows the most promise for recommending related articles in Wikipedia. We will investigate this hypothesis in future research.

To ensure reproducibility, we have made the data and source code of our study available at: https://github.com/wikimedia/citolytics.

7. REFERENCES

[1] Beel, J. et al. 2015. Research-paper recommender systems: a literature survey. Int. Journal on Digital Libraries. (2015).

[2] Beel, J. et al. 2016. Towards reproducibility in recommender-systems research. User Modeling and User-Adapted Interaction (UMAI). 26, (2016).

[3] Bellomi, F. and Bonato, R. 2005. Network Analysis for Wikipedia. Proc. of Wikimania. (2005).

[4] Bornmann, L. and Mutz, R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology. 66, 11 (2015).

[5] Cohen, D. et al. 2007. Lucene and Juru at Trec 2007 : 1-Million Queries Track. TREC 2007 (2007).

[6] Eto, M. 2013. Evaluations of context-based co-citation searching. Scientometrics. 94, 2 (2013).

[7] Garfield, E. 1964. Science Citation Index - a New Dimension in Indexing. Science. 144, 3619 (1964).

[8] Gipp, B. 2014. Citation-based Plagiarism Detection – Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis. Springer Vieweg Research.

[9] Gipp, B. et al. 2014. Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus. Journal of the American Society for Information Science and Technology. 65, 2 (2014).

[10] Gipp, B. et al. 2011. Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag. Proc. of 11th ACM/IEEE-CS Joint Conf. on Digital Libraries (JCDL'11) (2011).

[11] Gipp, B. et al. 2009. Scienstein: A research paper recommender system. Proc. of the Int. Conf. on Emerging Trends in Computing (ICETiC'09). (2009).

[12] Gipp, B. and Beel, J. 2009. Citation Proximity Analysis (CPA)-A new approach for identifying related work based on Co-Citation Analysis. Proc. of the 12th Int. Conf. on Scientometrics and Informetrics (ISSI'09). 2, (2009).

[13] Gipp, B. and Meuschke, N. 2011. Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. Proc. of the 11th ACM Symp. on Document Engineering (2011).

[14] He, Q. et al. 2010. Context-aware citation recommendation. Proc. of the 19th Int. Conf. on World Wide Web. (2010).

[15] Hu, M. et al. 2007. Measuring Article Quality in Wikipedia. Proc. of the 16th ACM Conf. on Information and Knowledge Management - CIKM '07 (2007).

[16] Huang, W. et al. 2015. A Neural Probabilistic Model for Context Based Citation Recommendation. Proc. of the 29th AAAI Conf. on Artificial Intelligence. (2015).

[17] Huang, W. et al. 2012. Recommending Citations : Translating Papers into References. Proc. of the 21st ACM Int. Conf. on Information and Knowledge Management. (2012).

[18] Joachims, T. et al. 2005. Accurately interpreting clickthrough data as implicit feedback. Proc. of the 28th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. (2005).

[19] Jones, K.S. 1973. Index term weighting. Information Storage and Retrieval.

[20] Konchady, M. 2008. Building Search Applications: Lucene, Lingpipe, and Gate.

[21] Leich, M. et al. 2013. Applying Stratosphere for Big Data Analytics. BTW (2013).

[22] Liu, S. and Chen, C. 2011. The effects of co-citation proximity on co-citation analysis. Proc. of ISSI. (2011).

[23] Marshakova, I. 1973. System of document connections based on references. Scientific and Technical Information Serial of VINITI. 6, (1973).

[24] McNee, S.M. et al. 2002. On the Recommending of Citations for Research Papers. Proc. of the 2002 ACM Conf. on Computer Supported Cooperative Work (2002).

[25] Ollivier, Y. and Senellart, P. 2007. Finding Related Pages Using Green Measures : An Illustration with Wikipedia. Proc. of AAAI. (2007).

[26] Page, L. et al. 1998. The PageRank Citation Ranking. World Wide Web Internet And Web Information Systems. 54, (1998).

[27] Paranjape, A. et al. 2015. Improving Website Hyperlink Structure Using Server Logs. arXiv preprint arXiv:1512.07258. (2015).

[28] Research: Wikipedia clickstream: 2015. http://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream. Accessed: 2015-05-27.

[29] Rubens, N. 2006. The Application of Fuzzy Logic to the Construction of the Ranking Function of Information Retrieval Systems. arXiv preprint cs/0610039. 10, (2006).

[30] Salton, G. et al. 1975. A Vector Space Model for Automatic Indexing. Communications of the ACM. 18, (1975).

[31] Small, H. 1973. A New Measure of the Relationship Two Documents. Journal of the American Society for Information Science. 24, (1973).

[32] Strohman, T. et al. 2007. Recommending Citations for Academic Papers. Proc. of the 30th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. (2007).

[33] Tran, N. et al. 2009. Enriching PubMed Related Article Search with Sentence Level. AMIA Annu. Symp. Proc. 2009, (2009).

[34] Wikipedia Clickstream: Getting Started: http://ewulczyn.github.io/Wikipedia_Clickstream_Getting_Started/. Accessed: 2015-05-25.

[35] Wikipedia:Manual of Style/Layout: 2014. https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Layout#See_also_section. Accessed: 2015-04-15.

[36] Woodruff, A. et al. 2000. Enhancing a Digital Book with a Reading Recommender. Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (2000).

[37] Wulczyn, E. and Taraborelli, D. 2015. Wikipedia Clickstream. figshare. (2015).