Preprint from https://www.gipp.com/pub/

M. Schwarzer, C. Breitinger, M. Schubotz, N. Meuschke, and B. Gipp. Citolytics - a link-based recommender system for wikipedia. In t.b.d, editor, *Proceedings of the 11th ACM Conference on Recommender systems (RecSys 2017)*, Aug. 2017

Citolytics - A Link-based Recommender System for Wikipedia

Malte Schwarzer[1], Corinna Breitinger[2], Moritz Schubotz[2], Norman Meuschke[2], Bela Gipp[2]
[1]National Institute of Informatics, Tokyo, Japan & Technical University Berlin, Germany
[2]University of Konstanz, Germany
ms@mieo.de,{first.last}@uni-konstanz.de

ABSTRACT

We present Citolytics - a novel link-based recommendation system for Wikipedia articles. In a preliminary study, Citolytics achieved promising results compared to the widely used text-based approach of Apache Lucene's MoreLikeThis (MLT). In this demo paper, we describe how we plan to integrate Citolytics into the Wikipedia infrastructure by using Elasticsearch and Apache Flink to serve recommendations for Wikipedia articles. Additionally, we propose a large-scale online evaluation design using the Wikipedia Android app. Working with Wikipedia data has several unique advantages. First, the availability of a very large user sample contributes to statistically significant results. Second, the openness of Wikipedia's architecture allows making our source code and evaluation data public, thus benefiting other researchers. If link-based recommendations show promise in our online evaluation, a deployment of the presented system within Wikipedia would have a far-reaching impact on Wikipedia's more than 30 million users.

CCS CONCEPTS

• **General and reference** → **Evaluation**; • **Information systems** → **Recommender systems**; *Wikis*; Evaluation of retrieval results;

KEYWORDS

Recommender System; Link-based; Wikipedia; Co-Citation Proximity Analysis; Online Evaluation

1 INTRODUCTION

Recommender systems (RS) are a crucial filtering and discovery tool to manage the vast and continuously increasing volume of items available in digital libraries and on the Web. However, RS are inherently difficult to evaluate: User-studies are expensive to conduct and are thus often small-scale. Offline evaluations, on the other hand, have been criticized for not accurately representing true user satisfaction [2]. Hence, our aim is to conduct a large-scale online evaluation for Citolytics, an open source RS framework designed for recommending related Wikipedia articles.

Working with Wikipedia data offers several benefits: (1) *Openness*. The corpus is publicly available and both source code and collected evaluation data can be published. This contributes to reproducibility. (2) *Size*. Wikipedia's very large user base contributes to statistical significance. (3) *Diversity*. Several million articles in 296 languages cover various topics. (4) *Transferability*. Wikis are widely used. This research can impact other domains and all websites using the MediaWiki[1] software.

In this work, we expand on our previous research on link-based RS [3], in which we conducted two independent large-scale *offline* evaluations using Wikipedia. In these previous evaluations, we investigated whether the two citation-based similarity measures Co-Citation (Co-Cit) and Co-Citation Proximity Analysis (CPA), the later of which Bela Gipp proposed in prior research [1], could be applied to the links in Wikipedia articles. We benchmarked these citation-based measures against the TF-IDF approach of MLT[2], which is currently being used by Wikipedia.

Both Co-Cit and CPA consider articles that are co-cited by other articles, or in the case of Wikipedia, co-linked. For example, if an article links to two other articles, article A and article B, then articles A and B are considered semantically related. The more frequently two articles are co-linked, the more likely they are semantically related. CPA additionally takes into account link proximity, i.e. the in-text placement of links. The closer two articles are co-linked in a text, e.g. sentence level vs. paragraph level, the more likely they are highly related. Relatedness is measured using the Co-Citation Proximity Index (CPI) [1].

The outcome of our initial study suggested that CPA and MLT performed similarly well in offline evaluations, while an additional manual analysis favored CPA. While MLT performed well in identifying narrowly similar articles that shared similar words and structure, CPA was better able to identify topically related information, such as information on the city of a certain university, or other technical universities in the region. However, the small scale of our previous manual analyses, and the shortcomings of offline evaluations, prevented us from arriving at definitive conclusions. Hence, we aim to conduct a large-scale online evaluation to arrive at a more conclusive judgment on the effectiveness of link-based recommendations.

2 SYSTEM OVERVIEW

To integrate Citolytics with Wikipedia, we must modify or access four components from the Wikipedia system as shown in Fig. 1:

(1) Wikipedia's Android app. The mobile app acts as the frontend in which recommendations are presented to users. The recommendations are shown under the heading "Read more" at the bottom of each article page. By default, three recommendations with their corresponding preview images are presented (Fig. 2). A screen capture video is available on YouTube[3]. The evaluation will

RecSys '17, , August 27–31, 2017, Como, Italy.

ACM ISBN 978-1-4503-4652-8/17/08.
https://doi.org/http://dx.doi.org/10.1145/3109859.3109981

[1]http://mediawiki.org/
[2]http://lucene.apache.org/
[3]https://youtu.be/gb09Z7PAlLU

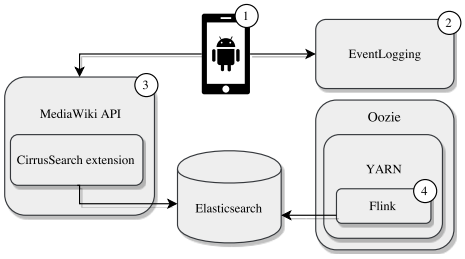

Figure 1: A Flink job scheduled by Oozie generates the recommendations and stores them in ES. An Android app accesses the recommendations via the MediaWiki API.

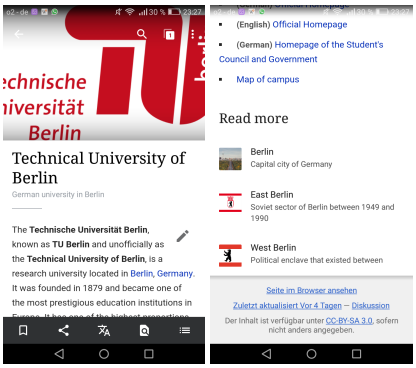

Figure 2: Screenshots of the recommender system within the Wikipedia Android app. Three recommendations are presented at the bottom of each article page.

be implemented as an A/B test: Users will randomly be assigned to one of two groups. While group A will use the current MLT implementation in Wikipedia, group B will receive the Citolytics recommendations.

(2) EventLogging. We must read from the EventLogging system, which logs user behavior in the Android app.

(3) MediaWiki. Recommendations must be made accessible via the MediaWiki API, which is the software that runs the Wikipedia website, and its CirrusSearch extension. Therefore, our modification must fulfill Wikipedia's scalability requirements.

(4) Oozie & YARN pipeline. The frequent generation of recommendations is ensured by an Apache Flink job[4] that is scheduled and executed in a pipeline based on Apache Oozie and YARN. The generated recommendations are written to Elasticsearch (ES), from where CirrusSearch reads the recommendations.

The Apache Flink job computes the article recommendations in batch fashion, whereby a Wikipedia XML dump is used as data source and an ES dump is the data sink, which is used to populate the recommendations to Wikipedia's ES index. The use of a big data processing framework like Flink is crucial for this task, since we must process all co-linked Wikipedia articles. For the English version of Wikipedia alone, this leads to intermediate results consisting of 11.8 billion records with a size of 0.8 TB. Fig. 3 shows the schema and descriptions of the involved Flink operators. While our demo setup uses a ten node cluster, Flink has been shown to be scalable to 1,000+ nodes.

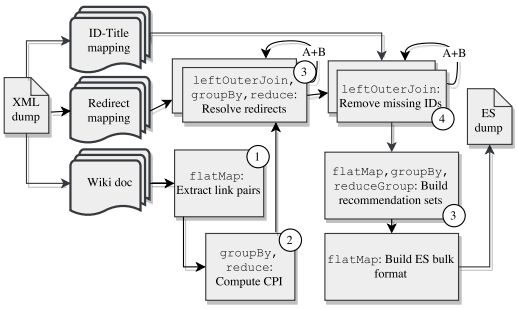

Figure 3: Schema of the Citolytics Flink job that generates the article recommendations. Operators marked with *A+B* are executed twice on each recommendation pair, once per recommendation direction.

The essential steps for the recommendation generation are: (1) extracting link pairs, (2) computing CPI, which is based on the original CPA concept [1], and (3) building recommendation sets. Additional steps, which mainly serve to tweak the recommendation performance include (4) redirection resolver, and (5) removal of missing IDs. **The Citolytics source code and documentation is available on GitHub**[5].

3 EVALUATION & CONTRIBUTION

The main contribution of this research will be its openly accessible and reproducible evaluation on a large scale. This online evaluation will be conducted using the data from Wikipedia's EventLogging system, which allows for detailed user tracking. Metrics such as click-through-rate, session length, and reading time of recommended articles will be collected.

4 CONCLUSION

With Citolytics we presented an open source RS for Wikipedia articles, and a proposal for its large-scale online evaluation. The ability to gather a large volume of user data will provide significant evidence to determine the best performing RS for the Wikipedia use case. Furthermore, the detailed categorization of Wikipedia articles will enable a granular evaluation of the strengths and weaknesses of each RS approach. The use of Wikipedia as a test system allows for the publication of the evaluation data, contributing to reproducibility. Finally, we would like to encourage other researchers to adapt our Citolytics framework[5] to their own RS needs.

REFERENCES

[1] B. Gipp and J. Beel. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. *ISSI '09: Proc. of the 12th Int. Conf. on Scientometrics and Informetrics*, 2:571–575, 2009.

[4] http://flink.apache.org/

[5] https://github.com/wikimedia/citolytics

[2] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012. DOI: 10.1007/s11257-011-9118-4.

[3] M. Schwarzer, M. Schubotz, N. Meuschke, C. Breitinger, V. Markl, and B. Gipp. Evaluating Link-based Recommendations for Wikipedia. *Proc. of the 16th ACM/IEEE Joint Conf. on Digital Libraries (JCDL '16)*:191–200, 2016. DOI: 10.1145/2910896.2910908.

APPENDIX

A SYSTEM DESCRIPTION

Citolytics is a link-based recommender system for Wikipedia articles. It has been designed for integration into the Wikipedia infrastructure, i.e. the MediaWiki software and the Wikipedia Android app, where the mobile app is used as the frontend to present the recommendations to the user. However, Citolytics is not strictly limited to the Wikipedia use case. Citolytics can easily be adapted to other websites that run a MediaWiki with the CirrusSearch extension.

B SETUP

The demo setup consists of three components. The MediaWiki environment must be set up in order to test the integration of Citolytics into the Wikipedia API. To generate new article recommendations, the Flink job is needed. However, we also make available for download[6] pre-generated recommendations using the simple English dump from Jan. 2017.

B.1 MediaWiki environment

A convenient way to set up the MediaWiki, which is the software that runs the Wikipedia website, is to use the portable development environment MediaWiki-Vagrant. MediaWiki-Vagrant consists of a set of configuration scripts for Vagrant and VirtualBox that automate the creation of a virtual machine that runs MediaWiki. A setup guide based on data from the simple English Wikipedia can be found on GitHub[7]. The guide shows how to create a basic MediaWiki setup, enable the search features from the CirrusSearch extension, install the Citolytics modifications and load demo data into the system.

For testing purposes, the recommendations can also be viewed directly from the MediaWiki search. To do this, you must use the query **citolytics:"Page title"** (with quotes) for Citolytics and **morelike:Page title** (without quotes) for MoreLikeThis, where the page title must be replaced with the page for which you want to retrieve recommendations.

B.2 Flink job

The Citolytics Flink job generates recommendations based on a Wikipedia XML dump. Apache Flink's jobs are made for distributed big data processing. Flink can be used stand-alone or integrated in a YARN environment. The computing cluster that runs the Flink job is required to provide at least 500 GB of temporary disk space. As a storage engine, we recommend HDFS, but all file systems supported by Flink are suitable. To run the Flink job, you must download the source code from GitHub[8], use Maven to build a JAR file, and submit the job to your Flink cluster. We explain the different job configurations on GitHub[9]. The generation of recommendations may take several hours depending on the size of the XML dump, the job configurations and the available computing power.

On our test system, a cluster of 10 IBM Power 730 (8231-E2B) servers, recommendation generation for the English Wikipedia took 3 hours. Each machine had 2x3.7 GHz POWER7 processors with 6 cores (12 cores 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.

B.3 Android app

The Android app serves as the frontend to present the recommendations. A recommender system is already a component of the official Wikipedia app, which can be downloaded from GooglePlay[10]. However, the official app provides only text-based MoreLikeThis recommendations. To enable the Citolytics recommendations, one must download the Citolytics fork of the Wikipedia Android app source from GitHub[11] and run the application from within an Android IDE (e.g. Android Studio).

Inside the app, the Wikipedia API endpoint must be changed to the endpoint of your MediaWiki setup. To do this, the developer settings must first be enabled. Developer settings can be enabled by tapping seven times on the Wikipedia circular icon that is located under **App settings → About the Wikipedia App**. After enabling, developer settings are available in the top right corner of the app settings screen.

C PRESENTERS

C. Breitinger is a doctoral researcher at the University of Konstanz. She received a Bachelor degree from the University of California, Berkeley and a Master degree from Linnaeus University in Sweden. Her interests lie in recommender systems, citation and link-based analyses, and the implications of blockchain-based applications.

Listing 1: Use the following `BibTeX` code to cite this article

```
@InProceedings{SchwarzerBSMG17,
  author    = {Malte Schwarzer and
    Corinna Breitinger and
    Moritz Schubotz and
    Norman Meuschke and
    Bela Gipp},
  title     = {Citolytics - A Link-based
      Recommender System for Wikipedia},
  booktitle = {Proceedings of the 11th ACM
      Conference on Recommender systems
    (RecSys  2017)},
  editor    = {t.b.d},
  year      = {2017},
  month     = {8}
}
```

[6] http://citolytics-demo.wmflabs.org/dumps/citolytics_simplewiki.json.gz
[7] https://github.com/mschwarzer/citolytics-demo
[8] https://github.com/wikimedia/citolytics
[9] https://github.com/wikimedia/citolytics/blob/master/support/flink-jobs/cirrussearch.md
[10] https://play.google.com/store/apps/details?id=org.wikipedia
[11] https://github.com/mschwarzer/apps-android-wikipedia