A Qualitative Evaluation of User Preference for Link-Based vs. Text-Based Recommendations of Wikipedia Articles

Malte Ostendorff[1(✉)], Corinna Breitinger[2], and Bela Gipp[2]

[1] Open Legal Data, Berlin, Germany
`mo@openlegaldata.io`
[2] University of Wuppertal, Wuppertal, Germany
`{breitinger,gipp}@uni-wuppertal.de`

Abstract. Literature recommendation systems (LRS) assist readers in the discovery of relevant content from the overwhelming amount of literature available. Despite the widespread adoption of LRS, there is a lack of research on the user-perceived recommendation characteristics for fundamentally different approaches to content-based literature recommendation. To complement existing quantitative studies on literature recommendation, we present qualitative study results that report on users' perceptions for two contrasting recommendation classes: (1) link-based recommendation represented by the Co-Citation Proximity (CPA) approach, and (2) text-based recommendation represented by Lucene's MoreLikeThis (MLT) algorithm. The empirical data analyzed in our study with twenty users and a diverse set of 40 Wikipedia articles indicate a noticeable difference between text- and link-based recommendation generation approaches along several key dimensions. The text-based MLT method receives higher satisfaction ratings in terms of user-perceived similarity of recommended articles. In contrast, the CPA approach receives higher satisfaction scores in terms of diversity and serendipity of recommendations. We conclude that users of literature recommendation systems can benefit most from hybrid approaches that combine both link- and text-based approaches, where the user's information needs and preferences should control the weighting for the approaches used. The optimal weighting of multiple approaches used in a hybrid recommendation system is highly dependent on a user's shifting needs.

Keywords: Information retrieval · Recommender systems · Human factors · Recommender evaluation · Wikipedia · Empirical studies

1 Introduction

The increasing volume of online literature has made recommendation systems an indispensable tool for readers. Over 50 million scientific publications are in circulation today [13], and approx. 3 million new publications are added annually

[15]. Encyclopedias such as Wikipedia are also subject to constant growth [39]. In the last decade, various recommendation approaches have been proposed for the literature recommendation use case. In a review of 185 publications, 96 different approaches for literature recommendation were identified. The majority of approaches (55%) continued to make use of content-based (CB) methods [4]. Only 18% of the surveyed recommendation approaches relied on collaborative filtering, and another 16% made use of graph-based recommendation methods, i.e. the analysis of citation networks, author, or venue networks [4]. A question that remains largely unexplored in today's literature is if these fundamentally different classes of recommendation algorithms are perceived differently by users. If a noticeable difference can be observed among users, across what dimensions do the end-users of such recommendation algorithms perceive that the approaches differ for a given recommendation use case? The majority of studies dedicated to evaluating LRS make use of offline evaluations using statistical accuracy metrics or error metrics without gathering any qualitative data from users in the wild [3]. More recently, additional metrics have been proposed to measure more dimensions of user-perceived quality for recommendations, e.g. novelty [11,29], diversity [22,41], serendipity [6,9,17,18], and overall satisfaction [14,27,43]. However, empirical user studies examining the perceived satisfaction with recommendations generated by different approaches remain rare. Given the emerging consensus on the importance of evaluating LRS from a user-centric perspective beyond accuracy alone [9], we identify a need for research to examine the user-perception of fundamentally different recommendation classes.

In this paper, we perform a qualitative study to examine user-perceived differences and thus highlight the benefits and drawbacks of two contrasting LRS applied to Wikipedia articles. We examine a text-based recommendation generation approach, represented by Lucene's MLT [1], and contrast this with a link-based approach, as implemented in the Citolytics recommendation engine [34], which uses co-citation proximity analysis (CPA) as a similarity measure [10]. Our study seeks to answer the following three research questions:

- RQ1: Is there a measurable difference in users' perception of the link-based approach compared to the text-based approach? If so, what difference do users perceive?
- RQ2: Do the approaches address different user information needs? If so, which user needs are best addressed by which approach?
- RQ3: Does one approach show better performance for certain topical categories or article characteristics?

Finally, we discuss how the evaluated recommendation approaches could be adapted in a hybrid system depending on the information needs of a user.

2 Background

2.1 Link-Based Similarity Measure

Co-citation proximity analysis (CPA; [10]) determines the similarity among articles by comparing the patterns of shared citations or links to other works within

the full text of a document. The underlying concept of CPA originates from the Library Science field, where it takes inspiration from the co-citation (CoCit) measure introduced by Small [37]. Beyond co-citation, CPA additionally takes into account the positioning of links to determine the similarity of documents. When the links of co-linked articles appear in close proximity within the linking article, the co-linked articles are assumed to be more strongly related.

Schwarzer et al. [34] applied the concept of CPA to the outgoing links contained in Wikipedia articles. They introduced Citolytics as the first link-based recommendation system using the CPA measure and applied it to the Wikipedia article recommendation use case. To quantify the degree of relatedness of co-linked articles, CPA assigns a numeric value, the Co-Citation Proximity Index (CPI), to each pair of articles co-linked in one or more linking articles. Schwarzer et al. [35] derived a general-purpose CPI that is independent of the structural elements of academic papers, e.g., sections or journals, which were proposed in the original CPA concept. The general-purpose CPI is, therefore, more suitable for links found in the Wikipedia corpus. We define the co-link proximity $\delta_j(a, b)$ as the number of words between the links to article a and b in the article j. Equation 1 shows the CPI for article a being a recommendation for article b.

$$\text{CPI}(a, b) = \sum_{j=1}^{|D|} \delta_j(a, b)^{-\alpha} * log(\frac{|D| - n_a + 0.5}{n_a + 0.5}) \qquad (1)$$

The CPI consists of two components: First, the general co-link proximity of a and b which is the sum of all marker proximities $\delta_j(a, b)$ over all articles in the corpus D. The parameter α defines the non-linear weighting of the proximity δ. In general, co-links in close proximity should result in a higher CPI than co-links further apart. Thus, α must be greater or equal to zero. Moreover, the higher α, the closer the co-link proximity must be to influence the final CPI score. Prior to the user study, we conducted an offline evaluation similar to [35] and found that CPA achieves the best results with $\alpha = 0.9$.

The second component of CPI is a factor that defines the specificity of article a based on its in-links n_a. This factor is inspired by the Inverse Document Frequency of TF-IDF, whereby we adapted the weighting schema from Okapi BM25 [38]. Hence, we refer to the factor as Inverse Link Frequency (ILF). We introduced ILF to counteract the tendency of CPA to recommend more general Wikipedia articles, which we discovered in a manual analysis of CPA recommendations. ILF increases the recommendation specificity by penalizing articles with many in-links, which tend to cover broad topics.

2.2   Text-Based Similarity Measure

The text based *MoreLikeThis* (MLT) similarity measure from Elasticsearch [7] (based on Apache Lucence) differs fundamentally from the CPA approach. Instead of links, MLT relies entirely on the terms present in the article text to determine similarity. Using a Vector Space Model [33], MLT represents articles as sparse vectors in a space where each dimension corresponds to a separate index

term. Term Frequency-Inverse Document Frequency (TF-IDF) proposed by [16] defines the weight of these index terms. Accordingly, MLT considers two articles similar the more terms they share and the more specific these terms are. Thus, MLT-based article recommendations are more likely to cover similar topics when the topic is defined by specific terms that do not occur in other topics. MLT's simplicity and ability to find similar articles has made it popular for websites. For instance, MLT is currently used by Wikipedia's MediaWiki software, as part of its *CirrusSearch* extension [28], to recommend articles to its users.

2.3   Related Work

Previously, Schwarzer et al. [35] performed an offline evaluation using the English Wikipedia corpus. They examined the offline performance of two link-based approaches, namely CPA and the more coarse CoCit measure, in addition to the text-based MLT measure. Schwarzer et al. made use of two quasi-gold standards afforded by Wikipedia. First, they considered the manually curated 'see also' links found at the end of Wikipedia articles as a quasi-gold-standard, which they used to evaluate 779,716 articles. Second, they used historical Wikipedia clickstream data in an evaluation of an additional 2.57 million articles. The results of this large-scale offline evaluation showed that the more fine-grained CPA measure consistently outperformed CoCit. This finding has also been validated by the research community in other recommendation scenarios [21,26].

Interestingly, this offline evaluation indicated that MLT performed better in identifying articles featuring a more narrow topical similarity with their source article. In contrast, CPA was better suited for recommending a broader spectrum of related articles [35]. However, prior evaluations by Schwarzer et al., using both 'See also' links (found at the bottom of Wikipedia articles) and clickstream data, were purely data-centric offline evaluations. This prohibits gaining in-depth and user-centric insights into the users' perceived usefulness of the recommendations shown. For example, click-through rates are a misleading metric for article relevance because users will click on articles with sensational or surprising titles before realizing that the content is not valuable to them. Accuracy and error metrics alone are not a reliable predictor of a user's perceived quality of recommendations [5]. Only user studies and online evaluations can reliably assess the effectiveness of real-world recommendation scenarios.

Knijnenburg et al. [20] proposed a user-centered framework that explains how objective system aspects influence subjective user behavior. Their framework is extensively evaluated with four trials and two controlled experiments and attempts to shed light on the interactions of personal and situational characteristics. Additionally, they take into account system aspects to explain the perceived user experience for movie recommendations.

Pu et al. [32] developed a user-centric evaluation framework termed ResQue (Recommender system's quality of user experience) consisting of 32 questions and 15 constructs to define the essential qualities of an effective and satisfying recommender system, including the recommendation qualities of accuracy,

novelty, and diversity. Since their framework can be applied to article recommendations, we include several questions from the ResQue framework in our evaluation design (refer to Sect. 3.1 for details).

Despite the large size and popularity of Wikipedia, the potential of this corpus for evaluating LRS has thus far not been exploited by the research community. To the best of our knowledge, no prior work, aside from the initial offline study [35] and the work by [30], has made use of the Wikipedia corpus to evaluate the effectiveness of different recommendation approaches. The implications of studying the user-perceived recommendation effectiveness for Wikipedia articles may also be applicable to Wikimedia projects in a broader context, which tend to contain a high frequency of links.

3   Methodology

This section describes our study methodology and the criteria for selecting the Wikipedia articles used in our study. The Wikipedia encyclopedia is one of the most prominent open-access corpora among online reference literature. As of June 2021, the English Wikipedia contains approximately 6.3 million articles [39]. Its widespread use and accessibility motivated us to use the English Wikipedia to source the articles for our live user study. We consulted the same English Wikipedia corpus as in [35] with the pre-processing as in [34].

3.1   Study Design

Prior to our study, we created a sample of 40 seed articles covering a diverse spectrum of article types in Wikipedia. When selecting these seed articles, our aim was to achieve a diversity of topics, which nonetheless remained comprehensible to a general audience. To ensure comprehensibility, we excluded topics that would require expert knowledge to judge the relevance of recommendations, e.g., articles on mathematical theorems. Moreover, the seed articles featured diverse article characteristics, such as article length and article quality[1].

We distinguished seed articles into four categories. First, according to their *popularity* (measured by page views) into either niche or popular articles, and second, according to the content of the article into either *generic*, i.e., reference articles typical of encyclopedias, or *named entities*, i.e., politicians, celebrities, or locations. We choose popularity as a criterion because, on average, popular articles receive more in-links from other articles. Schwarzer et al. [35] found that the number of in-links affected the performance of the link-based CPA approach. Moreover, we expect study participants to be more familiar with popular topics compared to niche articles. Therefore users will be better able to verbalize their spontaneous information needs when examining a topic. The 'article type' categories were chosen to study the effect that articles about named entities may have on MLT. Names of entities tend to be more unique than terms in articles

---

[1] We judge the article quality using Wikipedia's vital article policy [40].

on generic topics. Therefore, we expect that specific names may affect MLT's performance. Likewise, due to the nature of Wikipedia articles linking to generic topics, they may appear in a broader context than links to named entities. Thus, CPA's performance may also be affected. These considerations resulted in four article categories: (A) niche generic articles, (B) popular generic articles, (C) niche named entities, and (D) popular named entities. Table 1 shows these four categories and the 40 seed articles selected for recommendation generation.

To perform our qualitative evaluation of user-perceived recommendation effectiveness, we recruited 20 participants. Participants were students and doctoral researchers from several universities in Berlin and the University of Konstanz. The average age of participants was 29 years. 65% of our participants said they spend more than an hour per month on Wikipedia, with the average being 4.6 h spent on Wikipedia.

Our study contained both qualitative and quantitative data collection components. The quantitative component was in the form of a written questionnaire. This questionnaire asked participants about each recommendation set separately and elicited responses on a 5-point Likert scale. Some questions were tailored to gain insights on the research questions we defined for our study. The remainder of the questions adhered to the ResQue framework for user-centric evaluation [32]. The qualitative data component was designed as a semi-structured interview. The interview contained open-ended questions that encouraged participants to verbally compare and contrast the two recommendation sets. The participants were also asked to describe their perceived satisfaction. Resulting from this mixed methods study design, we could use the findings from the qualitative interviews to interpret and validate the results from the quantitative questionnaires. All interviews were audio-recorded with the permission of our participants.

In the study, each participant was shown four Wikipedia articles, one at a time. For each article, two recommendation sets, each containing five recommended articles, were displayed. One set was generated using CPA, i.e., the link-based Citolytics implementation [34], while the other was generated using the MLT algorithm. Each set of four Wikipedia articles was shown to a total of two participants to enable checking for the presence of inter-rater agreement. Participants were aware that recommendation sets had been generated using different approaches, but they did not know the names of the approaches or the method behind the recommendations. We alternated the placement of the recommendation sets to avoid the recognition of one approach over the other and forming a potential bias based on placement. The seed Wikipedia articles were shown to participants via a tablet or a laptop. The participants were asked to read and scroll through the full article so that the exploration of the article's content was a natural as possible. We have made the complete questionnaire and the collected data publicly available on GitHub[2].

---

[2] https://github.com/malteos/wikipedia-article-recommendations.

Table 1. Overview of seed articles selected for the study.

#	Article (Quality (See footnote 1))	Words	#	Article (Quality (See footnote 1))	Words
A	*Niche generic topics*		*C*	*Niche named entities*	
1	Babylonian mathematics (B)	3,825	21	Mainau (S)	567
2	Water pollution in India (S)	1,697	22	Lake Constance (C)	7,079
3	Transport in Greater Tokyo (C)	3,046	23	Spandau (C)	599
4	History of United States cricket (S)	3,610	24	Appenzell (C)	2,667
5	Firefox for Android (C)	4,821	25	Michael Müller (politician) (Stub)	602
6	Chocolate syrup (Stub)	391	26	Olympiastadion (Berlin) (C)	3,360
7	Freshwater snail (C)	1,757	27	Theo Albrecht (S)	929
8	Touring car racing (S)	2,550	28	ARD (broadcaster) (S)	2,397
9	Mudflat (C)	787	29	Kaufland (Stub)	680
10	Philosophy of healthcare (B)	3,804	30	Sylt Air (Stub)	110
B	*Popular generic topics*		*D*	*Popular named entities*	
11	Fire (C)	4,297	31	Albert Einstein (GA)	15,071
12	Basketball (C)	11,172	32	Hillary Clinton (FA)	28,645
13	Mandarin Chinese (C)	698	33	Brad Pitt (FA)	9,955
14	Cancer (B)	16,300	34	New York City (B)	30,167
15	Vietnam War (C)	32,847	35	India (FA)	16,861
16	Cat (GA)	17,009	36	Elon Musk (C)	11,529
17	Earthquake (C)	7,541	37	Google (C)	16,216
18	Submarine (C)	11,968	38	Star Wars (B)	16,046
19	Rock music (C)	19,833	39	AC/DC (FA)	10,442
20	Wind power (GA)	15,761	40	FIFA World Cup (FA)	7,699

4   Results

In this section, we summarize and discuss the empirical data collected. First, we present the primary findings, in which we provide answers to the three research questions specified in the Introduction, and illustrate them with participants' quotes. Second, we discuss secondary findings that arose from coding the participants' responses, which go beyond the research questions we set out to answer.

4.1   Primary Findings

Our study found several differences in reader's perception of the link-based approach compared to the text-based approach. A notable difference could be identified especially in the perceived degree of 'similarity' of the recommendations. Participants were significantly more likely to agree with the statement 'the recommendations are more similar to each other' (see 1.6 in Fig. 1) for the MLT approach. 73% of responses 'agreed' or 'strongly agreed' (58 out of 80 responses) with this statement, compared to only 36% of the responses for the CPA approach (29 out of 80). Keep in mind each of the 40 seed articles was examined by two participants resulting in 80 responses in total. A question about whether the articles being recommended 'matched with the content' of the source article (see 1.1) was answered with a similar preference, with a significantly higher portion of
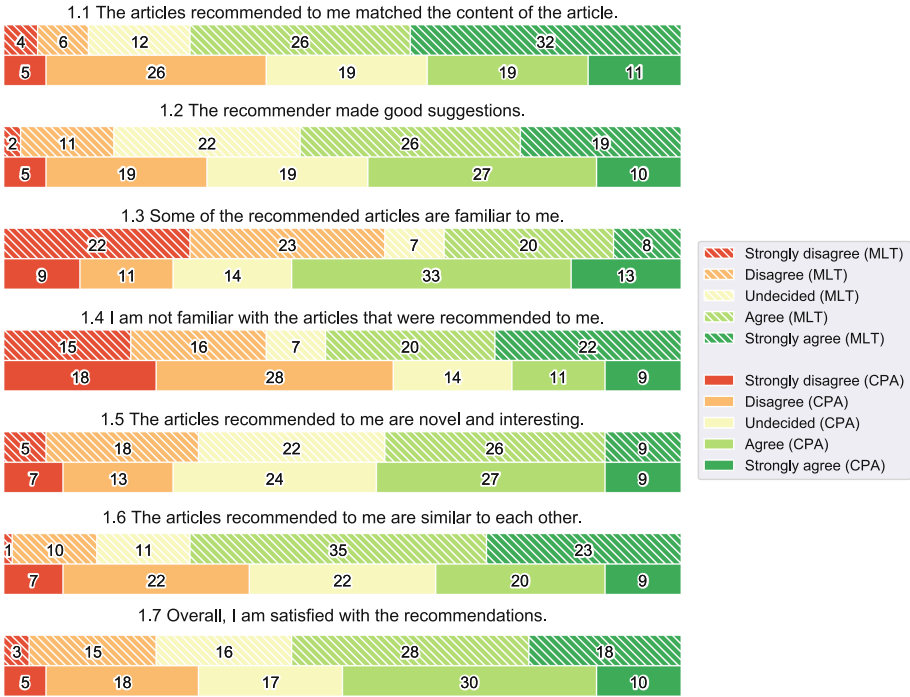
Fig. 1. Responses for MLT (dashed) and CPA (solid) on a 5-point Likert scale.

the responses indicating 'strongly agree' or 'agree' for the MLT approach (73%) and only 38% of responses choosing the same response for the CPA approach.

Overall, users perceived recommendations of CPA as more familiar (see 1.3). They felt less familiar (1.4) with the recommendations made by MLT. We found that this difference was observed by nearly all participants and can be attributed to how MLT considers textual similarity. In general, MLT focuses on overlapping terms, while CPA utilizes the co-occurrence of links. The quantitative results in [35] already suggested that this leads to diverging recommendations.

Perceived Difference between CPA and MLT. The participants observed that the methodological difference between the approaches affected their recommendations. In the questionnaire, participants expressed 48 times that the articles recommended by CPA are more diverse, i.e., less similar, compared to the seed article (Fig. 2a). MLT's recommendations were found to be diverse only 13 times. Regarding the similarity of recommendations, the outcome is the opposite.

The participants' answers also indicate the difference between MLT's and CPA's recommendations. Participant P20 explained that *"approach A [CPA] is more an overview of things and approach B [MLT] is focusing on concrete data or issues and regional areas"*. CPA providing an "overview of things" is not favorable for all participants as they describe different information needs.
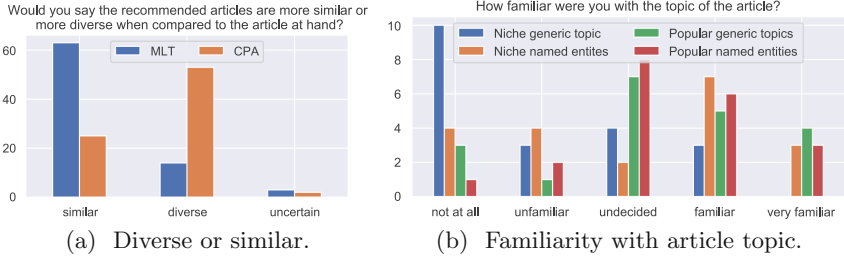
(a) Diverse or similar.

(b) Familiarity with article topic.

Fig. 2. Quantitative answers

For example, participant P20 prefers MLT's recommendation since *"it is bet-ter to focus on the details"*. Some participants attributed the recommendations' similarity (or diversity) to terms co-occurring in the title of the seed and recom-mended articles. For instance, participant P15 found MLT's recommendations for *Star Wars* to be more similar because *"Star Wars is always in the title [of MLT's recommendations]"*. Participant P17 also assumes a direct connection between the seed and MLT's recommendations *"I'd guess recommendations of A [MLT] are already contained as link in the source article"*.

Nonetheless, participants struggled to put the observed difference between MLT and CPA in words, although they noticed categorical differences in the recommendation sets. Participant P19 said *"I can see a difference but I don't know what the difference is"*. Similarly, participant P20 found that *"they [MLT and CPA] are both diverse to the same extent but within a different scope"*.

Concerning the overall relevancy of the recommendations, MLT outperformed CPA. In total, the participants agreed or strongly agreed 45 times that MLT made 'good suggestions' (see 1.2), whereas only 37 times the same was stated for CPA. Similarly, the overall satisfaction was slightly higher for MLT (46 times agree or strongly agree) compared to CPA (40 times; see 1.7).

Information Need. The participants are also aware of relevancy depending on their individual information need. When asked about the 'most relevant recom-mendation' the participants' answers contained the words 'depends' or 'depend-ing' ten-times. Participant P15 states that *"if I want a broader research I'd take B [CPA] but if I'd decide for more punctual research I would take A [MLT] because it is more likely to be around submarine and because in B [CPA] I also get background information"*. Similarly, participant P13 would click on a recom-mendation as follows: *"if you're looking for a specific class/type of snails then this [MLT] could be one, but if you're just looking to get an overview of aquatic animals, then probably you would click on the other approach [CPA]"*. In sum-mary, the participants agreed on CPA providing 'background information' that is useful to 'get an overview of a topic', while MLT's recommendations were perceived as 'more specific' and having a 'direct connection' to the seed article.

For articles on science and technology, the most commonly expressed infor-mation needs were understanding how a technology works or looking up a defini-

tion. For articles about individuals, participants expressed the need to find dates relating to an individual and to understand their contributions to society. For 'niche' topics, users were slightly more likely to state the desire to discover subcategories on a topic, which implies wishing to move from a broader overview to a more fine-grained and in-depth examination of the topic.

The subjectiveness is also reflected by the inter-rater agreement. The participants who reviewed the same articles had a Cohen's kappa of $\kappa = 0.14$ on average, which corresponds to slight agreement. The inter-rater agreement increases to a "fair agreement" (see [24]) when we move from a 5-point to a 3-point Likert scale, i.e., possible answers are 'agree', 'undecided', or 'disagree'. Low agreement indicates that the perception of recommendation highly depends on the individual's prior knowledge and information needs.

Article Characteristics. The article 'types', which we defined as described in the methodology section according to article popularity, length and breadth into the four categories 'popular generic', 'niche generic', 'popular named entities', and 'niche named entities' had no observable impact on user's preference for one recommendation approach over the other.

However, we found that the user-expressed information need, for example, the desire to identify related articles that were either more broadly related or were more specialized, did have a measurable impact on the user's preference for the recommendation method. For *popular generic articles* on science and technology, e.g., the article on wind power, the most frequently expressed information needs were understanding how a technology works or looking up definitions.

For articles in the categories 'popular generic' and 'niche generic', we could observe that the information needs expressed by our readers were more *broad*. For example, they wanted to find definitions for the topic at hand, more general information to understand a topic in its wider context, or examples of subcategories on a topic. There was no observable difference between the specified categories of information need for 'popular' vs. 'niche' generic articles.

Resulting from our initial classification of the 40 Wikipedia articles selected, the empirical questionnaire data showed that 'niche' entities were on average more familiar to the participants than we initially expected (Fig. 2b). This was especially the case for niche named entities, many of which were rated as being familiar to the participants. The reason for this may be that our participants were from Germany and were thus familiar with many of these articles, despite the articles reporting on regional German topics, e.g., Spandau, Mainau. On the other hand, users rated niche generic topics as being less familiar, which was in line with what we expected.

Furthermore, both popular generic topics and popular named entities were less often classified as 'unfamiliar' by the participants than they were classified as 'neutral or familiar'. Lastly, one notable finding is that Wikipedia listings, e.g., *List of rock genres* or *List of supermarket chains in Germany*, were found to be the most relevant recommendations in some cases. The CPA implementation [34] intentionally excludes Wikipedia listings from its recommendation sets. Thus, the implementation needs to be revised accordingly.
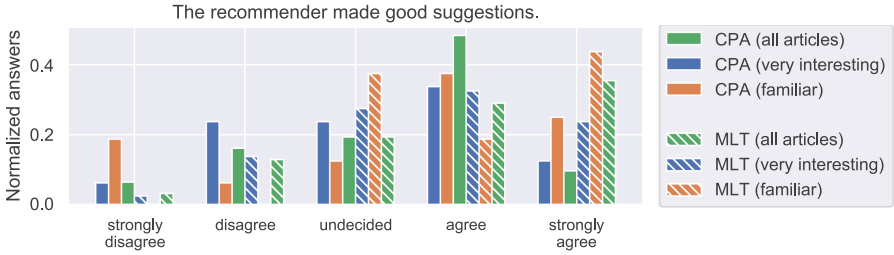
4.2   Secondary Findings

Fig. 3. User's satisfaction depending on interest and familiarity, i.e., for all articles or only the articles which are very interesting or familiar to the user.

Effect of User's Interest on Recommendations. Figure 3 shows that MLT outperforms CPA in recommendation satisfaction if participants 'strongly agreed' with the article topic being (a) interesting or (b) being at least 'familiar' to them. This is likely the case because users are more versed in judging the relevance of the text-based recommendations of the MLT approach if they already have more in-depth knowledge of a topic. For example, one participant observed regarding CPA's recommendations of *Renewable energy* for *Wind power* as the seed article: *"Renewable energy is least relevant because everybody knows something about it [Renewable energy]"*. Sinha and Swearingen [36] have already shown previous familiarity with an item as a confounding factor on a user 'liking' a recommendation. Interestingly, this trend was no longer observable in cases when participants only 'agreed' but without strong conviction that the articles were interesting or familiar to them. In these cases, the MLT and CPA approaches were seen as more equal, with the CPA approach taking a slight lead.

User-Based Preferences. Our findings confirm the subjectiveness of recommendation performance, since we observe user-based preferences. For instance, participant P2 only agreed or strongly agreed for CPA on MC 1.2 'The recommender made good suggestions' and 1.7 'Overall, I am satisfied with the recommendations' but never gave the same answers for MLT. However, participant P3 showed the opposite preference, i.e., only MLT made good suggestions according to P3. The remaining participants had more balanced preferences. In terms of MC 1.2 and 1.7, nine participants had a tendency to prefer MLT, while six participants preferred CPA, and five participants did not show any particular preference for one of the two recommendation approaches.

Perception of Novelty and Serendipity of Recommendations. To gain insights on the user-perceived novelty, we asked participants to rate the following statements: 'I am not familiar with the articles that were recommended to me' and

'The articles recommended to me are novel and interesting'. While novelty determines how unknown recommended items are to a user, serendipity is a measure of the extent to which recommendations positively surprise their users [9,18]. For instance, participant P19 answered *"approach A [MLT] shows me some topics connected with the article I read but with more special interest - they are about "Healthcare", and B [CPA] actually changes the whole topic. B [CPA] offers totally different topics."* CPA's recommendations are generally found to be more serendipitous. For the question regarding an 'unexpected recommendation among the recommendation set' CPA received a yes answer 41 times, compared to only 23 times for MLT. The perceived novelty also made participants click on recommendations. Among others, participant P1 explained that *"there are more [MLT] articles that I would personally click on, because they are new to me."* Similarly, participant P16 stated that they would *"click first on Star Wars canon, because I don't know what it is"*.

Trust and (Missing) Explanations. Although the questionnaire is not designed to investigate the participants' trust in the recommendations, many answers addressed this topic. When users were asked for the relevancy of recommendations, some participants expressed there "must be a connection" between the article at hand and a recommendation and that they just "do not know what is has to do with it". Others were even interested in topically irrelevant recommendations. For example, they expressed *"it interests me why this is important to the article I am reading"*. Similarly, a participant said they might click on a recommendation *"because I do not know what it has to do with [the seed article]"*. Such answers were more often found for CPA recommendations since they tend to be more broadly related than MLT's more narrow topical similarity. In some cases, there is no semantic relatedness. Yet, even then, participants often do not recognize a recommendation as irrelevant. Instead, they say it is their fault for not knowing how the recommendation is relevant to the seed. This behavior indicates a high level of trust from the participants placed in the recommender system.

5   Discussion

The experimental results demonstrate that MLT and CPA differ in their ability to satisfy specific user information needs. Furthermore, our study participants were capable of perceiving a systematic difference between the two approaches.

CPA was found to provide an 'overview of things' with recommendations more likely to be unfamiliar to the participants and less likely to match with the content of the seed article. In contrast, MLT was found to 'focus on the details'. Participants also felt that MLT's recommendations matched the content of the seed article more often. At the same time, participants perceive CPA's recommendations as more diverse, while MLT's recommendations are more similar to each other. So CPA and MLT, being conceptually different approaches and relying on different data sources, lead to unique differences in how the recommendations were perceived. In terms of the overall satisfaction with recommendations,

most participants expressed a preference for MLT over CPA. MLT is based on TF-IDF and, therefore, its recommendations are centered around specific terms (e.g., P15: *"Star Wars is always in the title"*). In contrast, CPA relies on the co-occurrence of links. According to CPA, two articles are considered related when they are mentioned in the same context. Our results show that this leads to more distantly related recommendations, which do not necessarily share the same terminology. Given that the participants experience the two recommendation approaches differently, a hybrid text- and link combination, depending on the context, is preferable (as demonstrated in [8,19,31]).

Moreover, the differently perceived recommendations show the shortcoming of the notion of similarity. Both approaches, CPA and MLT, were developed to retrieve semantically similar documents, which they indeed do [10,16]. However, their recommendations are 'similar' within different scopes. A recommended article that provides an 'overview' can be considered similar to the seed article. Equivalently, a 'detailed' recommendation can also be similar to the seed but in a different context. Our qualitative interview data could show how users perceive these two similarity measures differently. These findings are aligned with [2], which found that text similarity inherits different dimensions.

We also found that either CPA's or MLT's recommendations are liked or disliked depending on the individual participant preferences. Some participants even expressed a consistent preference for one method over the other. However, a strict preference was the exception. We could also not identify any direct relation between the user or article characteristics and the preference for one method. At this point, more user data as in a user-based recommender system would be needed to tailor the recommendations to the user's profile. Purely content-based approaches such as CPA and MLT lack this ability [4,12,25]. The only option would be to allow users to select their preferred recommendation approach through the user interface depending on their information need.

The participants' answers also revealed a trust in the quality of the recommendations that was not always justified. Participants would assume a connection between the seed article and the recommended article just because it was recommended by the system. Instead of holding the recommender system accountable for non-relevant recommendations, participants found themselves responsible for not understanding a recommendations relevance. To not disappoint this trust, recommender systems should provide explanations that help users understand why a particular item is recommended. Also, explanations would help users to understand connections between seed and recommendations. Explainable recommendations are a subject of active research [23,42]. However, most research focuses on user-based approaches, while content-based approaches like CPA or MLT could also benefit from explanations.

Despite the insights of our qualitative study to elicit user's perceived differences in recommendation approach performance, the nature of our evaluation has several shortcomings. With 20 participants, the study is limited in size. Consequently, our quantitative data points suggest a difference that is not statistically significant. Large-scale offline evaluations (e.g., [35]) are more likely to

produce statistically significant results. For this reason, and for not requiring participants, such offline evaluations are more commonly used in recommender system research [4]. But offline evaluations only provide insights in terms of performance measures. Our study shows that this can be an issue. When consulting only our quantitative data, one could assume that MLT and CPA are comparable in some aspects since their average scores are similarly high. The discrepancies between CPA and MLT only become evident when analyzing the written and oral explanations of live users. This highlights that recommender system research should not purely rely on offline evaluations [3].

Moreover, we acknowledge that recommendations of Wikipedia articles differ from recommendations of other literature types. It is thus uncertain whether our findings relating to encyclopedic recommendations in Wikipedia can be directly transferred to other domains. Recent advancements in recommender system research are focused on neural-based approaches, which may lead to the belief that the examined methods, MLT and CPA, are dated. This, however, is not the case since they are still used in practice, as Wikipedia's MLT deployment shows (Sect. 2.2), and the intuition of CPA is the basis for neural approaches like in Virtual Citation Proximity [30].

6   Conclusion

We elicit the user-perceived differences in performance for two well-known recommendation approaches. With the text-based MLT and the link-based CPA, we evaluate complementary content-based recommender system implementations. In a study with 20 participants, we collect qualitative and quantitative feedback on recommendations for 40 diverse Wikipedia articles.

Our results show that users are generally more satisfied with the recommendations generated by text-based MLT, whereas CPA's recommendation are perceived as more novel and diverse. The methodological difference of CPA and MLT, i.e., being based on either text or links, is reflected in their recommendations and noticed by the participants. Depending on information needs or user-based preferences, this leads to one recommendation approach being preferred over the other. Thus, we suggest combining both approaches in a hybrid system, since they both address different information needs.

As a result of the insights gained from our study, we plan to continue research on a hybrid approach tailored to the recommendation of literature, which accounts for diverse information needs. Moreover, we will investigate how content-based features can be utilized to provide explanation such that users can understand why a certain item is recommend to them. Lastly, we make our questionnaires, participants' answers, and code publicly available (See footnote 2).

References

1. Apache Lucence - More Like This. https://lucene.apache.org/core/7_2_0/queries/org/apache/lucene/queries/mlt/MoreLikeThis.html. Accessed 16 June 2021
2. Bär, D., Zesch, T., Gurevych, I.: A reflective view on text similarity. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011, pp. 515–520. Association for Computational Linguistics, Hissar (2011). https://www.aclweb.org/anthology/R11-1071
3. Beel, J., Breitinger, C., Langer, S., Lommatzsch, A., Gipp, B.: Towards reproducibility in recommender-systems research. User Model. User-Adap. Inter. **26**(1), 69–101 (2016). https://doi.org/10.1007/s11257-016-9174-x
4. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. Int. J. Digit. Libr. **17**(4), 305–338 (2015). https://doi.org/10.1007/s00799-015-0156-0
5. Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A.V., Turrin, R.: Looking for "Good" recommendations: a comparative evaluation of recommender systems. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011. LNCS, vol. 6948, pp. 152–168. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23765-2_11
6. De Gemmis, M., Lops, P., Semeraro, G., Musto, C.: An investigation on the serendipity problem in recommender systems. Inf. Process. Manag. **51**(5), 695–717 (2015). https://doi.org/10.1016/j.ipm.2015.06.008
7. ElasticSearch - More like this query. https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html. Accessed 16 June 2021
8. Färber, M., Sampath, A.: HybridCite: a hybrid model for context-aware citation recommendation. arXiv (2020). http://arxiv.org/abs/2002.06406
9. Ge, M., Delgado-Battenfeld, C., Jannach, D.: Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10, pp. 257–260. Association for Computing Machinery, New York (2010). https://doi.org/10.1145/1864708.1864761
10. Gipp, B., Beel, J.: Citation Proximity Analysis (CPA) – a new approach for identifying related work based on co-citation analysis. In: Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI '09), 2 July, pp. 571–575 (2009). https://ag-gipp.github.io/bib/preprints/gipp09a.pdf
11. Gravino, P., Monechi, B., Loreto, V.: Towards novelty-driven recommender systems. C R Phys. **20**(4), 371–379 (2019). https://doi.org/10.1016/j.crhy.2019.05.014
12. Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: Recommender Systems - An Introduction, 1st edn. Cambridge University Press, London (2010)
13. Jinha, A.E.: Article 50 million: an estimate of the number of scholarly articles in existence. Learn. Publish. **23**(3), 258–263 (2010). https://doi.org/10.1087/20100308
14. Joachims, T., Granka, L., Pan, B.: Accurately interpreting clickthrough data as implicit feedback. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2005). https://doi.org/10.1145/1076034.1076063
15. Johnson, R., Watkinson, A., Mabe, M.: The STM report: an overview of scientific and scholarly publishing. Int. Assoc. Sci. Tech. Med. Publish. **5**(October), 212 (2018)

16. Jones, K.S.: Index term weighting. Inf. Storage Retrieval **9**(11), 619–633 (1973). https://doi.org/10.1016/0020-0271(73)90043-0

17. Kaminskas, M., Bridge, D.: Measuring surprise in recommender systems. In: Rec-Sys REDD 2014: International Workshop on Recommender Systems Evaluation: Dimensions and Design, vol. 69, pp. 2–7 (2014)

18. Kaminskas, M., Bridge, D.: Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Trans. Interact. Intell. Syst. **7**(1), 1–42 (2017). https://doi.org/10.1145/2926720

19. Kanakia, A., Shen, Z., Eide, D., Wang, K.: A scalable hybrid research paper recommender system for microsoft academic. In: The World Wide Web Conference (WWW 2019), pp. 2893–2899. ACM Press, New York (2019). https://doi.org/10.1145/3308558.3313700

20. Knijnenburg, B.P., Willemsen, M.C., Gantner, Z., Soncu, H., Newell, C.: Explaining the user experience of recommender systems. User Model. User-Adap. Inter. **22**(4–5), 441–504 (2012). https://doi.org/10.1007/s11257-011-9118-4

21. Knoth, P., Khadka, A.: Can we do better than co-citations? - Bringing citation proximity analysis from idea to practice in research article recommendation. In: 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, Tokyo, Japan (2017)

22. Kunaver, M., Požrl, T.: Diversity in recommender systems - a survey. Knowl.-Based Syst. **123**, 154–162 (2017). https://doi.org/10.1016/j.knosys.2017.02.009

23. Kunkel, J., Donkers, T., Michael, L., Barbu, C.M., Ziegler, J.: Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In: Conference on Human Factors in Computing Systems - Proceedings (2019). https://doi.org/10.1145/3290605.3300717

24. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977). http://www.jstor.org/stable/2529310

25. Lenhart, P., Herzog, D.: Combining content-based and collaborative filtering for personalized sports news recommendations. In: Proceedings of the 3rd Workshop on New Trends in Content-Based RecommenderSystems (CBRecSys '16) at RecSys'16, vol. 1673, pp. 3–10 (2016)

26. Liu, S., Chen, C.: The effects of co-citation proximity on co-citation analysis. In: Proceedings of the 13th Conference of the International Society for Scientometrics and Informetrics (ISSI 2011) (2011)

27. Maksai, A., Garcin, F., Faltings, B.: Predicting online performance of news recommender systems through richer evaluation metrics. In: Proceedings of the 9th ACM Conference on Recommender Systems, pp. 179–186. ACM, New York (2015). https://doi.org/10.1145/2792838.2800184

28. Media Wiki Extension - CirrusSearch. https://www.mediawiki.org/wiki/Extension:CirrusSearch. Accessed 16 June 2021

29. Mendoza, M., Torres, N.: Evaluating content novelty in recommender systems. J. Intell. Inf. Syst. **54**(2), 297–316 (2019). https://doi.org/10.1007/s10844-019-00548-x

30. Molloy, P., Beel, J., Aizawa, A.: Virtual Citation Proximity (VCP): learning a hypothetical in-text citation-proximity metric for uncited documents. In: Proceedings of the 8th International Workshop on Mining Scientific Publications, pp. 1–8 (2020). https://doi.org/10.31219/osf.io/t5aqf

31. Ostendorff, M., Ash, E., Ruas, T., Gipp, B., Moreno-Schneider, J., Rehm, G.: Evaluating document representations for content-based legal literature recommendations. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, ICAIL '21, pp. 109–118. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3462757.3466073

32. Pu, P., Chen, L., Hu, R.: A user-centric evaluation framework for recommender systems. In: Proceedings of the Fifth ACM Conference on Recommender Systems - RecSys '11, p. 157. ACM Press, New York (2011). https://doi.org/10.1145/2043932.2043962

33. Salton, G., Wong, A., Yang, C.S.: Vector space model for automatic indexing. Information retrieval and language processing. Commun. ACM **18**(11), 613–620 (1975)

34. Schwarzer, M., Breitinger, C., Schubotz, M., Meuschke, N., Gipp, B.: Citolytics: a link-based recommender system for wikipedia. In: Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17, pp. 360–361. ACM, New York (2017). https://doi.org/10.1145/3109859.3109981

35. Schwarzer, M., Schubotz, M., Meuschke, N., Breitinger, C., Markl, V., Gipp, B.: Evaluating Link-based recommendations for wikipedia. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL '16, pp. 191–200. ACM Press, New York (2016). https://doi.org/10.1145/2910896.2910908

36. Sinha, R., Swearingen, K.: The role of transparency in recommender systems. In: CHI '02 Extended Abstracts on Human Factors in Computing Systems - CHI '02, p. 830. ACM Press, New York (2002)

37. Small, H.: Co-citation in the scientific literature: a new measure of the relationship between two documents. J. Am. Soc. Inf. Sci. **24**(4), 265–269 (1973). https://doi.org/10.1002/asi.4630240406

38. Sparck Jones, K., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments, Part 2. Inf. Process. Manag. **36**(6), 809–840 (2000). https://doi.org/10.1016/S0306-4573(00)00015-7

39. Wikipedia - Size of Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. Accessed 16 June 2021

40. Wikipedia - Vital articles. https://en.wikipedia.org/wiki/Wikipedia:Vital_articles. Accessed 16 June 2021

41. Yu, C., Lakshmanan, L., Amer-Yahia, S.: It takes variety to make a world: diversification in recommender systems. In: Edbt, pp. 368–378 (2009). https://doi.org/10.1145/1516360.1516404

42. Zhang, Y., Chen, X.: Explainable recommendation: a survey and new perspectives. Now Publish. (2020). https://doi.org/10.1561/9781680836592

43. Zhao, Q., Harper, F.M., Adomavicius, G., Konstan, J.A.: Explicit or implicit feedback? Engagement or satisfaction? a field experiment on machine-learning-based recommender systems. In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC '18, pp. 1331–1340. Association for Computing Machinery, New York (2018). https://doi.org/10.1145/3167132.3167275