information

MDPI

Article

Document Recommendations and Feedback Collection Analysis within the Slovenian Open-Access Infrastructure

Mladen Borovič *, Marko Ferme, Janez Brezovnik, Sandi Majninger, Klemen Kac and Milan Ojsteršek

Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia; marko.ferme@um.si (M.F.); janez.brezovnik@um.si (J.B.); sandi.majninger@um.si (S.M.); klemen.kac@um.si (K.K.); milan.ojstersek@um.si (M.O.)

* Correspondence: mladen.borovic@um.si; Tel.: +386-2-220-74-60

Received: 24 September 2020; Accepted: 21 October 2020; Published: 23 October 2020

Abstract: This paper presents a hybrid document recommender system intended for use in digital libraries and institutional repositories that are part of the Slovenian Open Access Infrastructure. The recommender system provides recommendations of similar documents across different digital libraries and institutional repositories with the aim to connect researchers and improve collaboration efforts. The hybrid recommender system makes use of document processing techniques, document metadata, and the similarity ranking function BM25 to provide content-based recommendations as a primary method. It also uses collaborative-filtering methods as a secondary method in a cascade hybrid recommendation technique. We also provide a real-world data feedback collection analysis for our hybrid recommender system on an academic digital repository in order to be able to identify suitable time-frames for direct feedback collection during the year.

Keywords: hybrid recommender systems; feedback collection; digital libraries; information retrieval; real-world data; open-access

1. Introduction

Recommender systems are a part of everyday experience on the web, especially while using online stores and search engines. The main objective of these systems is to provide the user with relevant and interesting content. In digital repositories, the obvious task for a recommender system is to provide recommendations to relevant documents. Digital repositories are usually used by students, researchers, and other interested parties, with an objective to research a certain topic and broaden their knowledge in that domain. A recommender system can be very helpful in achieving that, since it helps discover relevant documents, while the user does not need to browse and review a large amount of documents.

Recommender systems in academic digital repositories are becoming prominent as the number of produced academic documents in electronic format grows. There are many types of documents present in academic digital repositories, including, but not limited to, undergraduate theses, postgraduate theses (master's theses and doctoral theses), journal articles, conference articles, workbooks, study books, manuals, collections of problems, course slides, and other teaching and research materials. In Slovenia, universities, colleges, other higher education institutions, and research institutions have joined efforts to form the Slovenian Open-Access Infrastructure where documents from all partners would be publicly available. Naturally, this also provides a framework for recommender systems as it is possible to recommend documents between different institutions. Another positive side effect of this

is that researchers from different institutions that are in the same field of expertise can see the work of their colleagues more transparently, encouraging cooperation between them. With this goal in mind, a recommender system for the Slovenian Open-Access Infrastructure was designed as a part of the infrastructure to support the goals of the nationwide project. The novelty of this recommender system is that it is currently the only recommender system in Slovenia that includes all Slovenian universities and their electronic publications. In practice, over 200,000 electronic publications originating from any of the Slovenian universities can be recommended using our system.

This paper presents a cascade type hybrid recommender system which is implemented in the Slovenian Open-Access Infrastructure with the aim to serve relevant document recommendations across all digital libraries and institutional repositories which are currently included in the infrastructure. The second section briefly reviews related work. The third section presents the current state of the Slovenian Open Access Infrastructure. The inner workings and the architecture of our recommender system are presented in the fourth section. In the fifth section, we give details on the feedback collection analysis for our implemented hybrid recommender system using the digital repositories established within the Slovenian Open Access Infrastructure. The sixth section contains conclusions and ideas for further work.

2. Related Work

Document recommender systems can be applied in many practical scenarios. Specifically, for the scenario of document recommendations where the documents are news, Reference [1] demonstrates the use of recommendations for job postings, in Reference [2], cloud computing was used for recommendations and Reference [3] demonstrates a semantic web approach to recommending news. Many document recommender systems have been extensively covered by the research field especially for use with news. References [4,5] provide a survey of news recommendation systems. In [6], fuzzy logic is used to recommend news using content-based methods. Rich feedback is used to recommend news to users in [7], while Reference [8] compares information retrieval algorithms in news recommendation scenarios. In some cases, semantic approaches such as Wordnet are used to aid in semantic recommendations [9,10].

Research paper recommender systems are also prominent when it comes to document recommendations [11]. A tag-based research paper recommender system framework is presented in [12], and a similar tag-based approach was used in [13]. A collaborative filtering approach using contexts was used to recommend research papers in [14]. An extensive comparison of offline and online evaluation approaches of research paper recommender systems is presented in [15]. Specifically for digital repositories, several recommender systems have been developed. In [16], keyphrases were used as a basis for research paper recommendations and, in [17], a social bookmarking service CiteULike was used for recommendations. A recommender system specifically tailored for advising research publications as a part of digital libraries in a university environment was presented in [18]. Another study [19] introduces a Recommendation-as-a-Service (RaaS) platform used for recommendations in academia and its integration into the reference manager JabRef [20]. Similarly, CORE Recommender [21] was developed specifically for use in digital libraries and repositories. As shown in [22], such recommender systems have also been implemented in academic social networks, namely Mendeley.

When faced with researching, implementing, and maintaining recommender systems, challenges do occur. Some major challenges were outlined in [23]. These include data quality, the lack of appropriate data sets, choice of appropriate recommendation techniques, evaluation of recommendations, and even the number of recommended items. In addition to these challenges, we also encountered challenges while processing documents in the Slovenian language. Being a morphologically rich language, it is required to take different approaches to natural language processing when processing documents in Slovenian. Very little research has been done in recommending documents in the Slovenian language, mostly because there was very few structured

datasets of documents in Slovenian. With the introduction of the Slovenian Open Access Infrastructure [24], this has improved greatly due to the creation of a large structured dataset, containing over 200,000 documents [25]. It features segmented metadata consisting of titles, abstracts, keywords as well as full-texts and other document metadata. From it, other datasets of the Slovenian language have formed [26,27], which allows for further research options not only in the research of recommender systems, but also other tasks in information retrieval and natural language processing, specific to the Slovenian language.

3. Overview of the Slovenian Open Access Infrastructure

The Slovenian Open Access Infrastructure was established in 2013 and has since enabled the interested parties in Slovenia (researchers, students, companies, and the public) access to the intellectual production of Slovenian educational and research organizations. Simultaneously, it has enabled the researchers to fulfill the requirements for open access to publications from publicly financed research. Structurally (Figure 1), the infrastructure consists of a national portal OpenScience.si [28], institutional repositories for each of the four Slovenian universities (Digital Library of University of Maribor (DLUM) [29], Repository of University of Ljuljana (RUL) [30], Repository of University of Primorska (RUP) [31], Repository of University of Nova Gorica (RUNG) [32]), a repository for research institutions (Digital Repository of Slovenian Research Organizations (DiRROS) [33]), and a repository for colleges and higher education institutions (ReVIS) [34]).

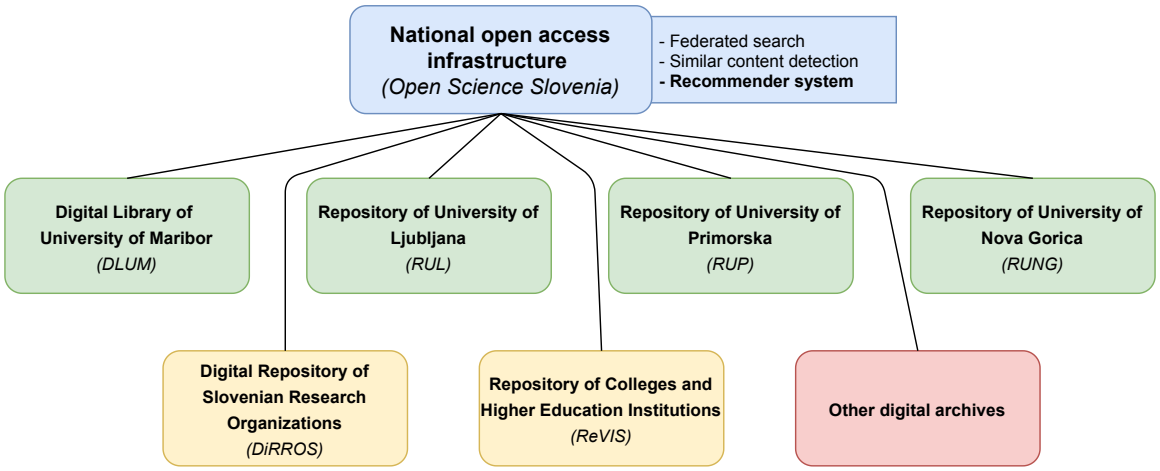

Figure 1. Structure of the Slovenian Open-Access Infrastructure.

The infrastructure also aggregates metadata from other digital archives such as videolectures.net [35], Social Science Data Archives [36], Digital Library of Slovenia [37], NUK Web Archive [38], and the Ministry of Defense Library and Information System [39]. The types of publications that are stored in the infrastructure include diploma, master's and doctoral theses, journal and conference articles, proceedings, datasets, scientific and technical reports, books, lecture materials, and videos of lectures. Since a great majority of publications are in Slovenian, a side product of this infrastructure was a large-scale corpus of full-text documents in the Slovenian language, covering several different domains of research. It also spawned some research datasets for use in linguistic studies [40,41]. More importantly, it currently represents the largest corpus of segmented texts in the Slovenian language, giving several options for research not only in linguistics but also in natural language processing. Due to interests for cooperation between the four universities and several research institutions in Slovenia, a recommender system was integrated in the infrastructure. The aim was to notify users about similar studies being done at different institutions through digital libraries and institutional repositories.

4. Document Recommendations

There are a few different approaches to recommendation in existence. The most common approaches are content-based and collaborative filtering [42,43]. Other approaches include demographic, utility-based, and knowledge-based techniques to recommendation. There is no optimal approach for every situation. Each approach has advantages and disadvantages in certain scenarios. While content-based filtering works well when a good description of an object is provided and when starting out with recommendations, collaborative filtering tends to provide more contextually appropriate recommendations once enough user feedback is provided. Hybrid systems aim to resolve the disadvantages of both approaches by combining them in different ways [44]. Several hybridization methods exist [45]. Weighted hybrids compute a score for a recommended item using outputs of all recommendation approaches available in the system. Switching hybrids employ a mechanism to switch between recommendation approaches. In this type of hybrid, approaches in the system are usually given priorities. If an approach with a higher priority cannot give a sufficient score, the recommender system switches to an approach with a lower priority as an attempt to provide a more recommendation with a more sufficient score. Mixed hybrids provide recommendations from different approaches at the same time. In cascade hybrids, one approach is used first to produce an initial set of recommended items; then, a second approach is used to fine-pick the most suitable items from that initial set, in order to provide a final recommendation.

Our recommender system is a cascade hybrid, incorporating content-based filtering as a primary recommendation technique and collaborative filtering as a secondary re-ranking method. It consists of three fundamental modules (Figure 2). The user activity log module provides the information on user activities such as view count, download count, document ratings, and document referrals. The document processing module ensures a unified feature representation of all documents in a triplet representation consisting of a title, keywords, and an abstract. Simultaneously, this module performs the calculation of BM25 values for each document pair, which forms a document index. The latter is a similarity matrix for all documents. Documents are periodically processed as new documents are added to the system daily. This way, the index is kept updated and the recommendations include new documents.

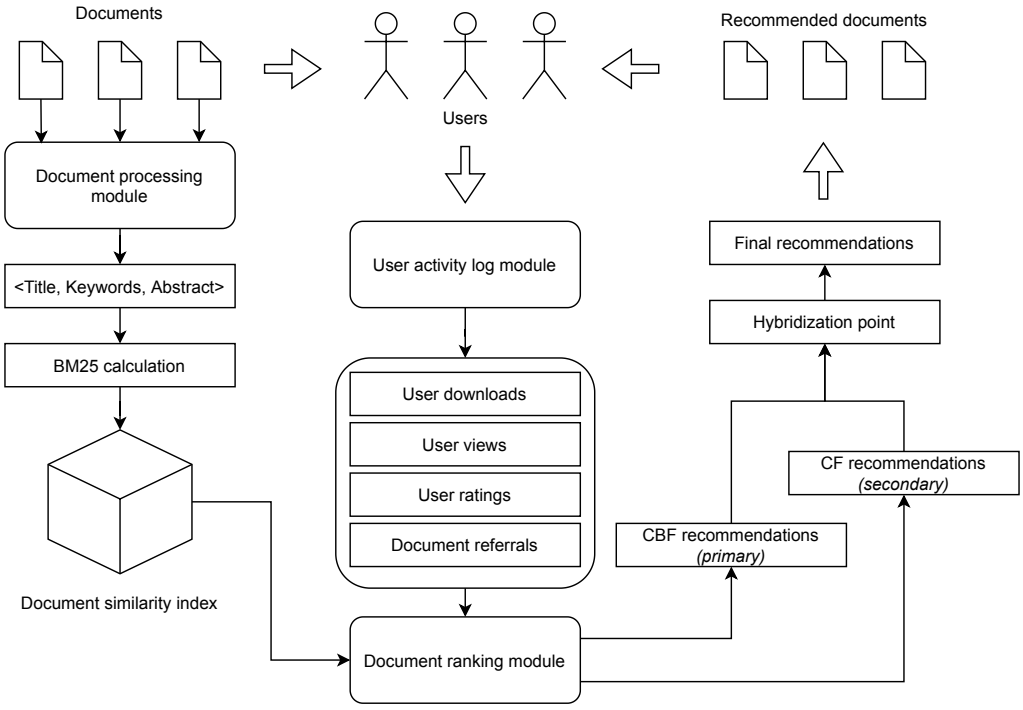

Figure 2. The architecture of our hybrid recommender system.

The user activity data and the calculated similarities between the documents are the input to the document ranking module, where similar documents are chosen depending on the document that is viewed by the user. This is also the hybridization point, where content-based filtering and collaborative filtering methods are applied in cascade to output the final list of recommendations, which is served to the end-user.

4.1. Processing Documents in Slovenian

A variety of different metadata were obtained from previous established repositories. These included information about authors, titles, keywords, abstracts, publishing year, and other bibliographic information. The metadata standards were different and included COMARC, MARC 21, and Dublin Core Metadata. We merged the different metadata schemes in our own metadata scheme to enable collection of as much metadata as possible. Our own metadata scheme consists of all metadata fields from the established standards with some extra fields for internal use. We use our metadata scheme to represent documents and use it with the recommender system as well as some other services within the Slovenian Open-Access Infrastructure.

For the recommender system, the documents are represented by titles, keywords, and abstracts. Most documents are in the Slovenian language; however, there are also documents in English, German, Italian, Croatian, and Hungarian. The documents that are not written in Slovenian have at least the abstract and keywords translated to Slovenian to conform with the publication and cataloguing rules. In the case of these documents, the available metadata in Slovenian are used with higher priority than the metadata in other languages. First, the most common words in the Slovenian language are removed from the text, since they do not contribute to semantic information. These are mainly conjunctions, prepositions, particles, and interjections; however, common verbs and nouns are also included. The common word list was built using word counts in documents. This is a periodic task, which is run each time after a recommendation index is updated. Additionally, we used lemmatization to help when dealing with conjugations and declensions in the text. Lemmatization is the process of determining the basic lexical form (i.e., lemma) to the words in a text. A very similar process to lemmatization is stemming. The main difference between lemmatization and stemming is that stemming does not convert the word into its dictionary form but simply cuts off the ending of the word. In text mining, lemmatization can be used to detect contexts of texts. It is used in our text processing step to group semantically similar words and to avoid the difficult process of grouping with declension and conjugation rules. Furthermore, n-grams for $N = [1, 2, 3, 4, 5]$ are generated and used with the tf-idf based ranking function BM25 to perform content-based filtering within our hybrid approach to recommendation.

4.2. Document Ranking

For document ranking, we used the BM25 ranking function [46] along with additional weights, which were obtained from document metadata and user activities. BM25 is a ranking function, which enables the ranking of documents by the similarity of terms that are contained within those documents. It is a family of functions, which differs by weighting schemes and parameter values. In general, tf and idf weights are used [47]. The term frequency (tf) is the occurrence count of a term t within a document d while the inverse document frequency (idf) is the importance of the term t in the given document collection D (Equation (1)). Composite nonlinear tf normalizations and the family of BM25 ranking functions have been used extensively in search engines to rank documents:

$$idf(t) = \log \frac{||D|| - n(t) + 0.5}{n(t) + 0.5} \tag{1}$$

$$s(d, Q) = \sum_{i=1}^{||Q||} idf(q_i) \cdot \frac{tf(q_i, d) \cdot (k_1 + 1)}{tf(q_i, d) + k_1 \cdot B}, \quad q_i \in Q, d \in D \tag{2}$$

$$B = 1 - b + b \cdot \frac{l_d}{avgdl} \tag{3}$$

It is a state-of-the-art $tf\text{-}idf$ based ranking function and has spawned many variants including BM25L, BM25+, BM25-adpt and BM25T [48,49], which bring improvements on very specific datasets. It has also been implemented in open source and commercial solutions such as Apache Lucene, Apache Solr, and Xapian as well as in Microsoft SQL Server and MySQL database implementations as a default full-text search solution. We decided to implement BM25 ourselves on a Microsoft SQL Server platform to have research options while studying parameters of the original ranking function and its variants, since commercial solutions do not allow enough customization. Another reason for this is that our documents are in the Slovenian language, for which only limited support exists in these open source and commercial solutions.

$||D||$ in Equation (1) is the length of the collection D and $n(t)$ is the number of documents which contain the term t. The BM25 value $s(d, q)$ depends on the weights tf and idf as well as parameters k_1 and b. A general BM25 calculation for a document d and a query q with terms q_i is given with Equation (2), where $||Q||$ is the size of the query Q given with the number of terms and B is a normalization factor (Equation (3)). In Equation (3), l_d is the length of document d and $avgdl$ is the average length of the document in the corpus D.

The parameter k_1 regulates the importance of the tf weight and the parameter b regulates the importance of document length. The values for these two parameters can be set using advanced optimization approaches, but usually values $k_1 \in [1.2, 2.0]$ and $b = 0.75$ are used [50]. Currently, we use empirically determined fixed values $k_1 = 1.2$ and $b = 0.75$, but further study of the corpus properties and parameter effects is underway. An automated adaptive technique of choosing the parameters using an optimization method such as in [51] is desired. Additionally, we are also working on including alternative weighting schemes such as tf*pdf [52] and $tf\text{-}id_uf$ [53].

4.3. Hybrid Approach to Recommendation

The input to our content-based filtering approach is a collection of metadata which describes the documents. A document feature is represented with a vector of terms obtained from titles, keywords, and abstracts. As we also have full-texts available, we empirically found that it is better to use semantically dense metadata rather than full-text due to two important disadvantages. Firstly, full-texts contain more terms which slows down the process of ranking similar documents. Secondly, semantically important contexts diminish even after applying pre-processing with stop-word lists and $tf\text{-}idf$ filtering. However, when compared to a simpler document feature assembled from titles, keywords and abstracts do not significantly improve recommendation results. We further enrich the document feature with metadata including document typology [54], issue year, authors, repository ID, and document language.

With all the metadata considered, we calculate a BM25 score based on the enriched document features. We also use the Jaro–Winkler distance [55,56], in order to define a document typology similarity. The Jaro–Winkler similarity is suitable when dealing with short strings and when the similarity between them should be greater if the two strings match from the beginning. First, the Jaro similarity is calculated by including the number of matching characters m and half the number of transpositions t between strings s_1 and s_2 and their respective lengths $||s_i||$ (Equation (4)). Then, the Jaro–Winkler similarity is calculated by including the common prefix length λ and a scaling factor $p = 0.1$ to adjust the value depending on the common prefix length (Equation (5)). In our situation, the document typologies are denoted with a short string of up to five characters (e.g., $\lambda = 5$). The first character of the typology defines the kind of document and the following characters define the variant of the document. Some examples of document types are provided in Table 1.

Table 1. Examples of document typologies and their metadata notation. Full typology is available in [54].

Document Typology (Notation)	Document Typology (Meaning)
1.01	Original scientific article
1.02	Review article
1.03	Short scientific article
1.04	Professional article
2.08	Doctoral dissertation
2.09	Master's thesis
2.11	Undergraduate thesis
2.23	Patent application
2.24	Patent
2.25	Other monographs and completed works

Using the Jaro-Winkler distance (Equation (6)), we compare the typologies of two documents in order to rank the documents with the similar typology higher. The final content-based filtering score (Equation (7)) is calculated as a product between the BM25 score on the document feature vector and the Jaro–Winkler similarity on the document typology:

$$sim_j(s_1, s_2) = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{4}$$

$$sim_{jw}(s_1, s_2) = sim_j(s_1, s_2) + \lambda p(1 - sim_j(s_1, s_2)) \tag{5}$$

$$d_{jw}(s_1, s_2) = 1 - sim_{jw}(s_1, s_2) \tag{6}$$

$$Score_{CBF} = BM25(d_A, d_B) \cdot d_{jw}(t_{d_A}, t_{d_B}) \tag{7}$$

Our collaborative filtering approach is collaborative in the sense that we use user interactions to re-rank the content-based filtering recommendations with the goal of improving recommendations. The input to our collaborative filtering approach is the user activity data regarding a document a_d. Views and download counts for documents are kept and regularly updated. The values for actions were set to 1 if a view occurs and 10 if a download occurs, meaning that a download action is as significant as 10 view actions (Equation (8)). A feedback value $f_{(a_d)}$ is calculated by summing all values of actions. Furthermore, we also store a similar feedback value for actions r_d on recommended documents $f_{(r_d)}$ to give higher weight to the documents which were interesting to end-users (Equation (9)). The values for boosts were set to 5 if a view on a recommended document occurs and 50 if a recommended document is downloaded. Action significance values for a_d and r_d were set empirically, with an idea in mind that a download is worth 10 times as significant as a view, and a recommended view is five times as significant as a regular view.

$$f_{a_d} = \sum_{i=1}^{||a_d||} a_{d,i} \qquad a_{d,i} = \begin{cases} 1 & \text{document view} \\ 10 & \text{document download} \end{cases} \tag{8}$$

$$f_{r_d} = \sum_{i=1}^{||r_d||} r_{d,i} \qquad r_{d,i} = \begin{cases} 5 & \text{document view} \\ 50 & \text{document download} \end{cases} \tag{9}$$

We can provide adaptive recommendations using actions from users by combining feedback values for actions and recommendations with the download rate h_d (Equation (10)), which is the ratio between downloads and views of a document. The logic is the same for the download rate of the recommended documents h_r, but only views and downloads on the recommended document are considered. The feedback value for actions on recommended documents makes the clicked

recommendations rank higher in the recommendation list. The final collaborative filtering score (Equation (11)) is calculated as a product of the document download rate and the sum of action feedback values on the document and actions on recommendations:

$$h_d = \frac{\text{downloads}(d)}{\text{views}(d)} \qquad h_r = \frac{\text{downloads}(d_r)}{\text{views}(d_r)} \tag{10}$$

$$Score_{CF} = f_{a_d} \cdot h_d + f_{r_d} \cdot h_r \tag{11}$$

With both approaches combined into a hybrid approach, we use recommendation strategies, which can be customized depending on the type or purpose of recommendations. Some recommendation strategies that we used in production are »latest + relevant«, »same repository + relevant« and »more from same authors«. These strategies can also be merged into a single strategy using priority factors. For example, a strategy »latest from same repository and from same authors« would first pick the latest documents and would then filter them according to their repository primarily and according to their authors secondarily.:

$$\tau_d = \delta^{\text{Year}_{\text{now}} - \text{Year}_d} \tag{12}$$

The workflow of our hybrid recommender system consists of four steps (Figure 3). First, the results from our content-based approach are obtained. Second, an exponential temporal decay mechanic (Equation (12)) is implemented to increase the ranks of recently published documents. The parameter δ controls the exponential temporal decay. The similarity score of the document is multiplied by the temporal decay and the recommendations in the results are re-ranked. Documents contained in the result set are then input into our collaborative filtering approach which re-ranks the results again. Currently, the output result length of our content-based approach is 25 documents. Finally, the list of recommendations is shortened to N documents for better presentation of the result on the web. In practice, we shorten the list to five documents.

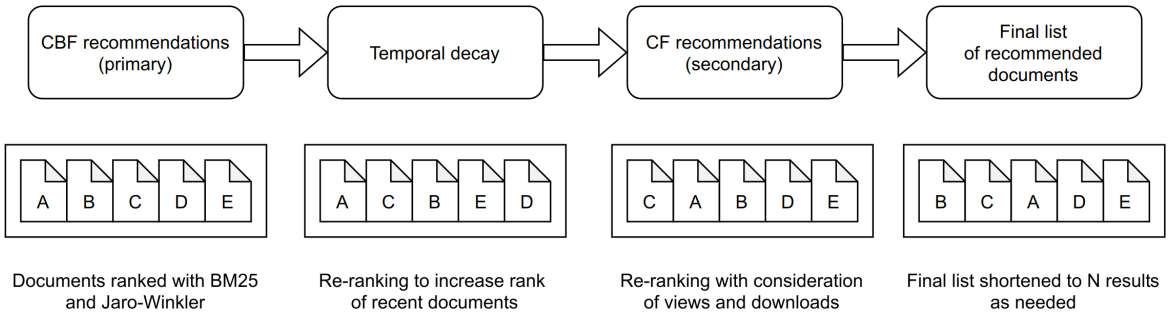

Figure 3. The workflow of our hybrid cascade approach to recommendations.

5. Feedback Collection Analysis

Collecting feedback from users is an important part of recommender systems design because it can directly influence the resulting recommendations. The overall user experience with regard to recommendations can be greatly improved if feedback is regularly collected from users. This can be done directly using surveys, questionnaires, and quick questions or indirectly by analyzing user activity. To achieve sufficient feedback, an appropriate time for feedback collection must be determined. The quality of feedback depends on the mood of the user, but, with careful planning, there is more chance that the user will be willing to give good quality feedback. Another perspective is to collect feedback at a certain time, where we are sure that users might be more inclined to express their opinions (e.g., a week after something changed) as they have had enough time to form an opinion. Furthermore, a good feedback collection approach can lead to an organized approach to evaluation of

recommender systems. With it, evaluation metrics can be better defined and used to measure the true performance of the recommender system.

We performed an analysis of time-frames during the year, when feedback collection would make sense within the Slovenian Open-Access Infrastructure. In our case, the recommendations are focused on documents and are meant to help students, academic staff, and researchers find more similar documents to their interest. The recommendations are therefore accessed as the users are using the recommender system, which is linked to different time-frames during the year. We found that several spikes in usage occur during the year and we tried to link them to specific events that occur in the academic year (e.g., thesis defenses, summer vacations, etc.).

We limited our data to data from four universities in Slovenia and their institutional repositories in the Slovenian Open-Access Infrastructure. University of Maribor was included with DLUM, University of Ljubljana with RUL, University of Primorska with RUP and University of Nova Gorica with RUNG.

All institutional repositories store view and download counts for documents. During this analysis, we treated viewed documents as mildly interesting and downloaded documents as very interesting. We did this because a download can occur only after the document is viewed; therefore, if a user downloaded the document, they must have viewed its detailed description with metadata and made a conscious decision that it is interesting enough for them to download it.

We encountered a major limitation with the accessibility of the traffic data on each institutional repository. DLUM was the only repository that we were able to get the data from, since other repositories opted not to be included in the analysis by their maintenance teams. Furthermore, the maintenance teams of DLUM, RUL, RUP, and RUNG decided to exclude all traffic tracking options on repositories after 2016. As for DLUM data that we were able to obtain, it was Google Analytics traffic data between January 2013 and December 2016. With all limitations considered, we performed an analysis using data only from DLUM (Figure 4). It proved to be a suitable institutional repository for this task, since it is the first university institutional repository in Slovenia, running since 2008 and serving as a basis for all other institutional repositories in the national open-access infrastructure.

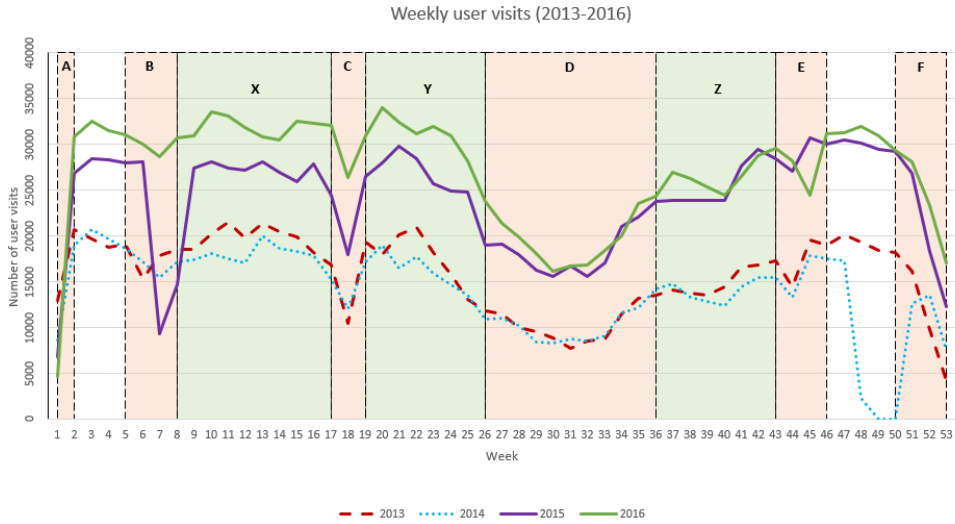

Figure 4. Weekly user visits to DLUM between January 2013 and December 2016.

In the data set time-frame of user activity between January 2013 and December 2016 (Figure 4), special events have occurred. In November 2014, DLUM saw a major update and was offline for two weeks (weeks 48 to 50) due to this. It was updated at this time because it had to run stable for most of the year, due to a regular influx of new theses. This influx annually reaches a peak in September and October (weeks 40 to 42), when the theses are catalogued by the librarians. It was decided to run

DLUM without interruption between March and November 2014 because most users during that time are students researching for their theses and researchers searching for related work for their articles.

An increase in weekly user visits can be observed in 2015. This increase seems to be attributed to the marketing efforts of the Slovenian Open-Access Infrastructure and the cross-repository recommendations; however, this cannot be confirmed due to the lack of traffic tracking capabilities on repositories RUL, RUP, and RUNG.

Furthermore, in 2016, we can observe another increase in weekly user visits, which lasts from January (week 1) to September (week 40). This unusual additional traffic was generated by students enrolled in pre-Bologna process study programs at the University of Maribor. These students had to complete and defend their theses by October 2016 as directed by the University of Maribor and were most likely collecting research on DLUM in order to achieve this. This reason holds, as the traffic increase stops in September 2016 (week 40).

By observing traffic fluctuation during the year, we found a decrease in weeks that correspond to holidays. This occurs in several time-frames which are visible in Figure 4 and denoted with letters:

- A—January; the first week of the year (consequence of New Year),
- B—February; weeks 7 and 8, around February 8th (national holiday "Prešeren Day"),
- C—April and May; week 18 and 19, starting around April 27th (national holiday "Day of uprising against occupation") and ending around May 1st (national holiday "International Workers' Day"),
- D—June, July and August; weeks 26 to 36, summer holiday season,
- E—October, November; weeks 44 and 45, around October 31st (national holiday "Reformation Day") and November 1st (national holiday "All Saint's Day"),
- F—December; weeks 50 to 53, around December 25th (national holiday "Christmas"), 26th (national holiday "Independence and Unity Day") and December 31st (national holiday "New Year's Eve").

We conclude that these time-frames are suitable for maintenance work on institutional repositories. Time-frames B, C, and E show the potential for smaller updates and minor changes, while time-frame D shows the potential for large-scale maintenance.

We also observed the peak traffic occurring between some before mentioned time-frames:

- X—weeks 9 and 17 (from February to April),
- Y—weeks 20 to 25 (from May to June),
- Z—weeks 37 to 43 (from August to October).

We conclude that these time-frames are suitable for feedback collection campaigns, surveys, and questionnaires. Namely, time-frames X and Y are more suitable for active user feedback collection (e.g., validation of recommended documents), since users are actively researching during that time. Time-frame Z is more suitable for general feedback collection (e.g., general surveys regarding user experience).

An extensive evaluation study of our recommender system is currently still underway as it requires successful collaboration of several institutions that maintain their own repositories. Several metrics for recommendation system evaluation exist. In general, there are two ways of evaluating any recommendation system: online and offline [15,57,58]. Offline evaluation makes use of preferably labelled data which is split into training and test sets. The recommendation system uses the training set ratings to try and predict the ratings in the test set. Actual users are not needed in this type of evaluation. This makes offline evaluation fast and easy to perform on a large amount of data. It can also be performed using many different datasets and with multiple different algorithms. The main disadvantage of this approach is that it cannot measure true user satisfaction.

In an online evaluation scenario, users interact with a running recommendation system and respond to it naturally, while feedback is being collected from them. Feedback is obtained by either asking the users directly or observing their actions. This approach measures true user satisfaction but can take a long time to set-up and run from beginning to end.

The choice of metrics differs depending on the approach of recommendation. Information retrieval metrics such as accuracy, recall, precision, and F-measure are usually considered preferable when evaluating content-based recommendation systems. Other metrics for this type of recommendation system include normalized discounted cumulative gain [59], rank-biased precision [60], and expected reciprocal rank [61]. Collaborative filtering recommendations are usually evaluated using approaches that measure novelty, serendipity, diversity, and coverage [62]. Currently, there are several different metrics [63] that can be used to evaluate recommendation systems. When dealing with hybrid recommendation systems, this must be carefully considered, since the type of hybridization can also affect the evaluation process, making it complex due to implementation in multiple stages.

6. Conclusions

In this article, we present a cascade hybrid recommender system implemented in institutional repositories that is part of the Slovenian National Open-Access Infrastructure. We outlined the recommender system architecture, document pre-processing, and ranking approaches. A feedback collection analysis has been presented on real-world data from one of our longest running repositories. With the analysis, we were able to identify different time-frames during the year where it is suitable to consider feedback collection on an academic digital repository. An extensive evaluation study is currently underway and we conclude that, for an extensive evaluation of our recommender system's contribution to knowledge exchange and spread across the Slovenian Open-Access Infrastructure, a unified framework should be developed in addition to institutional repository management processes regarding logging user activities and using traffic tracking scripts. Only with such an approach can a definitive contribution of the recommender system be confirmed and further researched. It would also allow the observation of any significant cooperation between institutions, as it is already suspected that the institutions in the two largest institutional repositories in the national open-access infrastructure be in accordance with the majority of research cooperation efforts in Slovenia.

Author Contributions: Conceptualization, M.B., M.F., and M.O.; methodology, M.B. and M.F.; software, M.B., M.F., and J.B.; validation, M.B., M.F., J.B., S.M., K.K., and M.O.; writing—original draft preparation, M.B. and M.O.; writing—review and editing, M.B., M.F., J.B., S.M., K.K., and M.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elsafty, A.; Riedl, M.; Biemann, C. Document-based Recommender System for Job Postings using Dense Representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 3, pp. 216–224. [CrossRef]
2. del Campo, J.V.; Pegueroles, J.; Hernández-Serrano, J.; Soriano, M. DocCloud: A document recommender system on cloud computing with plausible deniability. *Inf. Sci.* **2014**, *258*, 387–402. [CrossRef]
3. Cantador, I.; Bellogín, A.; Castells, P. News@hand: A Semantic Web Approach to Recommending News. In *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Proceedings of the AH 2008: Adaptive Hypermedia and Adaptive Web-Based Systems, Hannover, Germany, 29 July–1 August 2008*; Nejdl, W., Kay, J., Pu, P., Herder, E., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 279–283.
4. Karimi, M.; Jannach, D.; Jugovac, M. News recommender systems—Survey and roads ahead. *Inf. Process. Manag.* **2018**, *54*, 1203–1227. [CrossRef]
5. Borges, H.L.; Lorena, A.C. A Survey on Recommender Systems for News Data. In *Smart Information and Knowledge Management: Advances, Challenges, and Critical Issues*; Szczerbicki, E., Nguyen, N.T., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 129–151. [CrossRef]

6. Adnan, M.N.M.; Chowdury, M.R.; Taz, I.; Ahmed, T.; Rahman, R.M. Content based news recommendation system based on fuzzy logic. In Proceedings of the 2014 International Conference on Informatics, Electronics Vision (ICIEV), Dhaka, Bangladesh, 23–24 May 2014; pp. 1–6.
7. Ardissono, L.; Petrone, G.; Vigliaturo, F. News Recommender Based on Rich Feedback. In *International Conference on User Modeling, Adaptation, and Personalization, Proceedings of the UMAP 2015: User Modeling, Adaptation and Personalization, Dublin, Ireland, 29 June–3 July 2015*; Ricci, F., Bontcheva, K., Conlan, O., Lawless, S., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 331–336.
8. Bogers, T.; van den Bosch, A. Comparing and Evaluating Information Retrieval Algorithms for News Recommendation. In Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys '07), Minneapolis, MN, USA, 19–20 October 2007; pp. 141–144. [CrossRef]
9. Capelle, M.; Hogenboom, F.; Hogenboom, A.; Frasincar, F. Semantic News Recommendation Using Wordnet and Bing Similarities. In Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC '13), Coimbra, Portugal, 18–22 March 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 296–302. [CrossRef]
10. Capelle, M.; Moerland, M.; Hogenboom, F.; Frasincar, F.; Vandic, D. Bing-SF-IDF+: A Hybrid Semantics-Driven News Recommender. In Proceedings of the 30th Annual ACM Symposium on Applied Computing (SAC '15), Salamanca, Spain, 13–17 April 2017; pp. 732–739. [CrossRef]
11. Beel, J.; Gipp, B.; Langer, S.; Breitinger, C. Research-paper recommender systems: A literature survey. *Int. J. Digit. Libr.* **2016**, *17*, 305–338. [CrossRef]
12. Jomsri, P.; Sanguansintukul, S.; Choochaiwattana, W. A Framework for Tag-Based Research Paper Recommender System: An IR Approach. In Proceedings of the 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, Perth, WA, Australia, 20–23 April 2010; pp. 103–108.
13. Choochaiwattana, W. Usage of tagging for research paper recommendation. In Proceedings of the 2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), Chengdu, China, 20–22 August 2010; Volume 2, pp. V2-439–V2-442.
14. Winoto, P.; Tang, T.; McCalla, G. Contexts in a Paper Recommendation System with Collaborative Filtering. *Int. Rev. Res. Open Distance Learn.* **2012**, *13*, 56–75. [CrossRef]
15. Beel, J.; Langer, S. A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems. In *International Conference on Theory and Practice of Digital Libraries, Proceedings of the TPDL 2015: Research and Advanced Technology for Digital Libraries, Poznan, Poland, 14–18 September 2015*; Kapidakis, S., Mazurek, C., Werla, M., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 153–168.
16. Ferrara, F.; Pudota, N.; Tasso, C. A Keyphrase-Based Paper Recommender System. In *Digital Libraries and Archives*; Agosti, M., Esposito, F., Meghini, C., Orio, N., Eds.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 14–25.
17. Bogers, T.; van den Bosch, A. Recommending Scientific Articles Using Citeulike. In Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08), Lausanne, Switzerland, 23–25 October 2008; pp. 287–290. [CrossRef]
18. Porcel, C.; Moreno, J.; Herrera-Viedma, E. A multi-disciplinar recommender system to advice research resources in University Digital Libraries. *Expert Syst. Appl.* **2009**, *36*, 12520–12528. [CrossRef]
19. Beel, J.; Aizawa, A.; Breitinger, C.; Gipp, B. Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia. In Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Toronto, ON, Canada, 19–23 June 2017; pp. 1–2. [CrossRef]
20. Feyer, S.; Siebert, S.; Gipp, B.; Aizawa, A.; Beel, J. Integration of the Scientific Recommender System Mr. DLib into the Reference Manager JabRef. In *European Conference on Information Retrieval, Proceedings of the ECIR 2017: Advances in Information Retrieval, Aberdeen, UK, 8–13 April 2017*; Springer: Cham, Switzerland, 2017. [CrossRef]
21. Knoth, P.; Anastasiou, L.; Charalampous, A.; Cancellieri, M.; Pearce, S.; Pontika, N.; Bayer, V. Towards effective research recommender systems for repositories. *arXiv* **2017**, arXiv:1705.00578.
22. Vargas, S.; Hristakeva, M.; Jack, K. Mendeley: Recommendations for Researchers. In Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16), Boston, MA, USA, 15–19 September 2016; p. 365.

23. Beel, J.; Dinesh, S. Real-World Recommender Systems for Academia: The Pain and Gain in Building, Operating, and Researching them [Long Version]. *arXiv* **2017**, arXiv:1704.00156.
24. Ojsteršek, M.; Brezovnik, J.; Kotar, M.; Ferme, M.; Hrovat, G.; Bregant, A.; Borovič, M. Establishing of a Slovenian open access infrastructure: A technical point of view. *Program* **2014**, *48*, 394–412. [CrossRef]
25. OpenScience Slovenia Dataset. Available online: http://www.openscience.si/OpenData.aspx (accessed on 23 October 2020).
26. Erjavec, T.; Fišer, D.; Ljubešić, N.; Arhar Holdt, Š.; Bren, U.; Robnik Šikonja, M.; Udovič, B. Terminology Identification Dataset KAS-Term 1.0. Available online: https://www.clarin.si/repository/xmlui/handle/11356/1198 (accessed on 23 October 2020).
27. Erjavec, T.; Fišer, D.; Ljubešić, N.; Bitenc, M. Bilingual Terminology Extraction Dataset KAS-Biterm 1.0. Available online: https://www.clarin.si/repository/xmlui/handle/11356/1199 (accessed on 23 October 2020).
28. OpenScience Slovenia. Available online: https://www.openscience.si/ (accessed on 23 October 2020).
29. Digital Library of University of Maribor-DLUM. Available online: https://dk.um.si/info/index.php/eng (accessed on 23 October 2020).
30. Repository of the University of Ljubljana-RUL. Available online: https://repozitorij.uni-lj.si/info/index.php/eng (accessed on 23 October 2020).
31. Repository of the University of Primorska-RUP. Available online: https://repozitorij.upr/info/index.php/eng (accessed on 23 October 2020).
32. Repository of the University of Nova Gorica-RUNG. Available online: https://repozitorij.ung.si/info/index.php/eng (accessed on 23 October 2020).
33. Digital repository of Slovenian Research Organizations. Available online: https://dirros.openscience.si/info/index.php/eng (accessed on 23 October 2020).
34. Repository of Colleges and Higher Education Institutions-ReVIS. Available online: https://revis.openscience.si/info/index.php/eng (accessed on 23 October 2020).
35. Videolectures.net. Available online: https://videolectures.net (accessed on 23 October 2020).
36. Social Science Data Archives. Available online: https://www.adp.fdv.uni-lj.si/eng/ (accessed on 23 October 2020).
37. Digital Library of Slovenia. Available online: http://dlib.si/?=&language=eng (accessed on 23 October 2020).
38. NUK Web Archive. Available online: http://arhiv.nuk.uni-lj.si (accessed on 23 October 2020).
39. Ministry of Defence Library and Information System. Available online: https://dk.mors.si/info/index.php/en (accessed on 23 October 2020).
40. Jakubíček, M.; Fiser, D.; Suchomel, V. Terminology Extraction for Academic Slovene Using Sketch Engine. In Proceedings of the Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, Karlova Studanka, Czech Republic, 2–4 December 2016; Volume 10.
41. Ljubešić, N.; Fiser, D.; Erjavec, T. KAS-term: Extracting Slovene Terms from Doctoral Theses via Supervised Machine Learning. In *International Conference on Text, Speech, and Dialogue, Proceedings of the TSD 2019: Text, Speech, and Dialogue, Ljubljana, Slovenia, 11–13 September 2019*; Springer: Cham, Switzerland, 2019; pp. 115–126. [CrossRef]
42. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749. [CrossRef]
43. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl.-Based Syst.* **2013**, *46*, 109–132. [CrossRef]
44. Burke, R. Hybrid Recommender Systems: Survey and Experiments. *User Model. User-Adapt. Interact.* **2002**, *12*, 331–370. [CrossRef]
45. Burke, R. Hybrid Web Recommender Systems. In *The Adaptive Web: Methods and Strategies of Web Personalization*; Brusilovsky, P., Kobsa, A., Nejdl, W., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 377–408. [CrossRef]
46. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
47. Jones, K.; Walker, S.; Robertson, S. A probabilistic model of information retrieval: Development and comparative experiments: Part 2. *Inf. Process. Manag.* **2000**, *36*, 809–840. [CrossRef]
48. Géry, M.; Largeron, C. BM25t: A BM25 extension for focused information retrieval. *Knowl. Inf. Syst.* **2012**, *32*, 217–241. [CrossRef]

49. Trotman, A.; Puurula, A.; Burgess, B. Improvements to BM25 and Language Models Examined. In Proceedings of the 2014 Australasian Document Computing Symposium (ADCS '14), Melbourne, VIC, Australia, 27–28 November 2014; ACM: New York, NY, USA, 2014; pp. 58–65. [CrossRef]
50. Manning, C.D.; Raghavan, P.; Schütze, H. In *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008.
51. Bollegala, D.; Noman, N.; Iba, H. RankDE: Learning a Ranking Function for Information Retrieval Using Differential Evolution. In Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation (GECCO '11), Dublin, Ireland, 12–16 July 2011; pp. 1771–1778. [CrossRef]
52. Nguyen, K.; Shin, B.-J.; Yoo, S.J. Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information. In Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp), Hong Kong, China, 18–20 January 2016; pp. 223–230.
53. Beel, J.; Langer, S.; Gipp, B. TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections. In Proceedings of the iConference 2017, Wuhan, China, 22–25 March 2017; doi: 10.9776/17217. [CrossRef]
54. COBISS/IZUM, Typology of Documents/Works for Bibliography Management in COBISS. 2016. Available online: https://home.izum.si/COBISS/bibliografije/Tipologija_eng.pdf (accessed on 23 October 2020).
55. Jaro, M.A. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *J. Am. Stat. Assoc.* **1989**, *84*, 414–420. [CrossRef]
56. Winkler, W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Available online: https://files.eric.ed.gov/fulltext/ED325505.pdf (accessed on 23 October 2020).
57. Hernández del Olmo, F.; Gaudioso, E. Evaluation of recommender systems: A new approach. *Expert Syst. Appl.* **2008**, *35*, 790–804. [CrossRef]
58. Silveira, T.; Zhang, M.; Lin, X.; Liu, Y.; Ma, S. How good your recommender system is? A survey on evaluations in recommendation. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 813–831. [CrossRef]
59. Wang, Y.; Wang, L.; Li, Y.; He, D.; Liu, T.Y. A Theoretical Analysis of NDCG Type Ranking Measures. In *Conference on Learning Theory*; PMLR: Princeton, NJ, USA, 2013; Volume 30, pp. 25–54.
60. Moffat, A.; Zobel, J. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Trans. Inf. Syst.* **2008**, *27*. [CrossRef]
61. Chapelle, O.; Metlzer, D.; Zhang, Y.; Grinspan, P. Expected Reciprocal Rank for Graded Relevance. In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09), Hong Kong, China, 2–6 November 2009; Association for Computing Machinery: New York, NY, USA, 2009; pp. 621–630. [CrossRef]
62. Gunawardana, A.; Shani, G. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *J. Mach. Learn. Res.* **2009**, *10*, 2935–2962.
63. Shani, G.; Gunawardana, A. Evaluating Recommendation Systems. In *Recommender Systems Handbook*; Springer: Boston, MA, USA, 2011; pp. 257–297. [CrossRef]