Recommender Systems based on Scientific Publications: A Systematic Mapping

Felipe Ciacia de Mendonça[a], Isabela Gasparini[b], Rebeca Schroeder[c] and Avanilde Kemczinski[d]
UDESC - Santa Catarina State University, Joinville, SC, Brazil

Keywords: Recommender System, Scientific Publications, Systematic Mapping.

Abstract: Recommender Systems are intended to recommend items according users' preference, resulting in greater satisfaction to them. Among the objects of study that may be recommended are scientific articles from venues such as conferences and journals. However, there are still many challenges in this area, such as effective analysis of textual data as well as improvement of the recommendations produced. In this paper we investigate the state-of-the-art. For this purpose, we have applied the systematic mapping methodology (SM), considering 165 articles selected from the search string. Applying the inclusion criteria resulted in 78 articles, and applying the exclusion criteria resulted in 38 articles to answer the defined research questions. As result, it is possible to know which evaluation approaches, algorithms, and metrics are being used, as well as which databases are being studied for research in the area.

1 INTRODUCTION

Technology has brought many significant advances in our society, however, also brought some consequences. Among them, there was an increase in the amount of data in different application domains Tan et al. (2009). Such increase has led to a qualitative change in the methods of processing data, and there are different analyzes that can be performed to extract predictions, temporal analyzes, and other useful information to aid decision making Nassirtoussi et al. (2014). Understanding users's interests has become increasingly complex as a result of the growing mass of data Skiena (2017). This challenge has given rise to Recommender Systems (RSs), which, reduce the user's difficulty in finding items they want more effectively and faster Park et al. (2012).Thus, there is a reduction of information overload delivered to users through personalized information.

Among the types of data that need proper treatment there are textual data, which bring greater complexity in processing compared to numeric data Brunialti et al. (2015). In this case, complexity can be attributed to different interpretations, grammar, spelling, and even by languages. Examples of textual data are scientific publications from events such as conferences, workshops, symposiums, as well as journals. The Natural Language Processing (NLP) area enables this study, allowing the discovery of valuable information from publications through syntactic and semantic analysis of texts. NLP analyzes includes: Text Summary, Textual Linking, Prediction, Categorization, Topic Segmentation, Information Extraction, In-Text Sentiment Analysis, among others Skiena (2017). In this paper, the application domain is the scientific publications, in order to explore how researchers have been using RSs to recommend publications and venues based on textual data from scientific publication.

Some systematic bibliographic research developed for the subareas related to this study are highlighted such as: Machine Learning Malhotra (2015); Palaniappan et al. (2013); Portugal et al. (2017), Natural Language Processing (NLP) Brunialti et al. (2015); Nassirtoussi et al. (2014); Pons et al. (2016), and Recommender Systems (RSs) Champiri et al. (2015); Park et al. (2012). However, there are no systematic reviews or mappings for RSs based on scientific publications using textual analysis. Therefore, as a way of knowing the state-of-the-art developed for recommendations based on scientific publications, it is essential to carry out a systematic mapping on the subject. Thus, the aim of this paper is to identify the

[a] https://orcid.org/0000-0002-7732-448X
[b] https://orcid.org/0000-0002-8094-9261
[c] https://orcid.org/0000-0001-8882-3375
[d] https://orcid.org/0000-0001-7671-5457

Ciacia de Mendonça, F., Gasparini, I., Schroeder, R. and Kemczinski, A.
Recommender Systems based on Scientific Publications: A Systematic Mapping.
DOI: 10.5220/0009356307350742
In *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS 2020) - Volume 1*, pages 735-742
ISBN: 978-989-758-423-7

state of the art in RS based on scientific publications. To this end, a systematic mapping of the literature was conducted.

2 METHODOLOGY: SYSTEMATIC MAPPING PROCESS

A systematic mapping is a study that seeks to identify, evaluate, and interpret all available research relevant to a particular research question Kitchenham (2004). Therefore the objective of this Systematic Mapping of Literature (SML) looking for articles that have developed, evaluated or described Recommender Systems focusing on scientific publications, to suggest publications or scientific articles. To conduct this systematic mapping we applied the process proposed by Petersen et al. (2015).

This mapping aims to answer the Main Research Question: "How are Recommender Systems (RS) being used to assist in choosing publication venues?". To this end some secondary research questions (SRQs) have been defined in order to help answer the main question: SRQ1. From which countries are the institutions of the authors of the published studies? SRQ2. How old are the publication of the studies? SRQ3. Which are the type of publications (journals, conferences)? SRQ4. Which RS approaches are being used? SRQ5. Which databases were used? SRQ6. Which algorithm(s) are being used? and SRQ7. How is the recommendation process evaluated in terms of metrics?

Secondary questions aim to guide research to find current and relevant work. Each of the questions was answered by analyzing the articles resulting from the search performed.Based on the elaborated research questions, the search performed in the Scientific Search Engines (SSEs) used the Wazlawick (2017) method as a reference, which suggests investigating the technique itself that will be used and the target area of the research. The seach string used in each SSE is available online [1].

According to Buchinger et al. (2014), who conducted a quantitative analysis with 40 available Scientific Research Engines, the following mechanisms are relevant to the Computer Science area and are among the top 10 in their analysis: *ACM DL, IEEE Xplore, Science Direct, Springer Link*, and *Scopus*. Search strings tailored for each SSE are available online [1].

The number of articles returned for each SSE is shown in Table 1. The initial idea of the search string has been adapted for each SSE because a few results are different of the expected given that some SSEs automatically recognize plural words, and others do not have the same appeal.

Table 1: Number of Publications per SSE.

Scientific Search Engine	Nº of Publications
Springer Link	3594
Scopus	106
ACM Digital Library	34
IEEE Xplore	18
Science Direct	7
Total	**3759**

The selection of systematic mapping studies consists of applying a set of objective and subjective criteria (inclusion and exclusion) to be included or excluded from the classification. First, the objective criteria (OC) were applied, which were defined as follows:

Objective Criteria (OC): OC1. Publication Date: Any publication date; OC2. Type: Scientific Articles; OC3. Language: English only; OC4. Availability: Available for download; OC5. Access: free or available from our university; and OC6. Size: Full Papers (with 4 or more pages).

As a result, Springer was removed from this study because the filters applied returned more than 3000 articles, which made manual filtering difficult. Based on the 165 articles captured from the search engines *ACM DL*, *Science Direct*, *Scopus*, and *IEEE Xplore*, the inclusion (IC) and exclusion (EC) criteria were applied as follows: **Inclusion Criteria (IC):** IC1. Include articles that effectively address the research focus; IC2. Selection of primary works. **Exclusion Criteria (EC):** EC1. Derived articles (translations, extensions, etc); EC2. Studies that did not involve RSs with scientific articles; EC3. Duplicate articles; EC4. The article without an abstract;and EC5. Studies that could not be fully accessed.

Forty duplicate articles were found and 87 articles were rejected by the inclusion and exclusion criteria as shown in Figure 1. At the end of the selection of articles by applying the inclusion and exclusion criteria, the final result of 38 articles was obtained for the systematic mapping. The list of thirty-eight articles selected in this systematic mapping is available online [1].

From the thirty-eight resulting articles, their metadata was extracted and the articles were analyzed. Among the data that were analyzed are: year of publication, location of authors' institutions, and venues. After classifying these data, we began the analysis of the articles, based on reading, understanding and ex-

[1]https://bit.ly/2LonD40

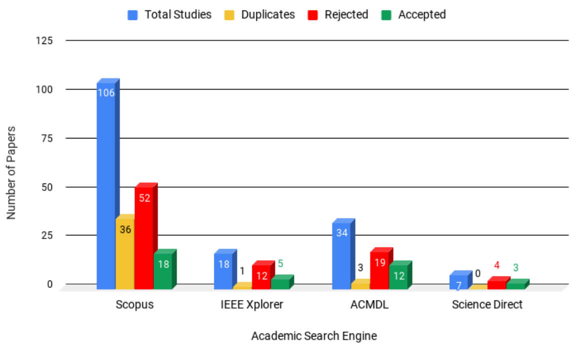

Figure 1: Number of articles for each scientific search engine (SSE).

tracting information that would be useful to answer the research questions. The information extracted includes: recommender system approaches used in each article, databases used to compose the system for each work, algorithms chosen for use in the system, and metrics used to evaluate the RS.

3 RESULTS

In this section, the results of the systematic mapping are presented, based on the data extracted from the 38 resulting articles, as well as the analyzes performed in the selected articles.

3.1 From Which Countries are the Institutions of the Authors of the Published Studies?

After performing data extraction, the 38 resulting articles were analyzed in order to answer the Main Research Question. However, this requires answering Secondary Research Questions (SRQ). According to Figure 2, it can be seen from which places are the institutions of the authors who have published on the subject of this mapping, thus answering SRQ1. The USA and China stand out from other countries in this segment, followed by some first world countries such as Italy, Germany and Australia.

3.2 How Old are the Publication of the Studies?

It is seen that the studies are very recent. Figure 3 shows a timeline of articles years, with the oldest being from 2008. This shows that this subject began to be studied 10 years ago, and has been growing gradually.

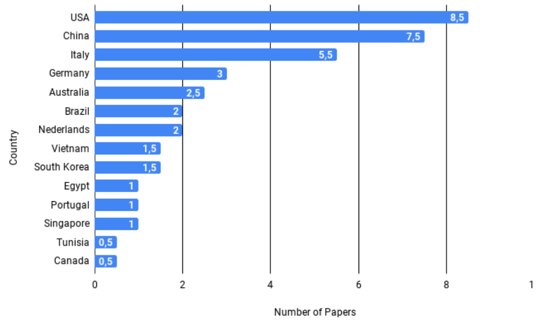

Figure 2: Number of articles by country.

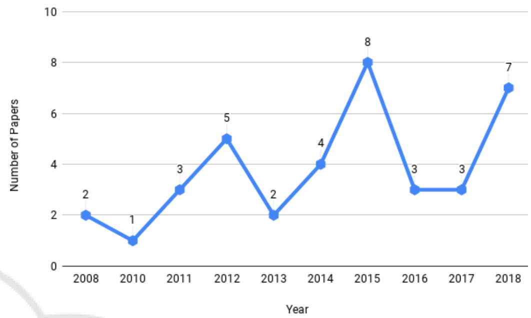

Figure 3: Timeline with years of publications.

3.3 Which are the Type of Publications?

In response to SRQ3, which concerns publication vehicles, the numbers show diverse events and journals. Among the journals, there were 10 different journals, with only one study each. In the tables 2 and 3 it can be seen that among the conferences, there were 28 articles published at events such as conferences or workshops, 24 of which were distinct, with only one of them standing out with more than one occurrence, the ACM Conference on Recommender Systems (RecSys) conference with five publications. It is noted that for these topics of study, the authors seek to publish more at events such as conferences and workshops rather than journals.

3.4 Which RS Approaches are Being Used?

Supported by the approaches of Taghavi et al. (2017), Adomavicius and Tuzhilin (2005), Burke (2002, 2007), Ibrahima and Younisb (2018), Ricci et al. (2015), and Jannach et al. (2010), we sought to classify the studies according to the Recommender Systems approaches that were the most used in order to respond to SRQ4. The use of the approach by models based on collaborative filtering had a total of seventeen papers (44.7%), while ten papers (26.3%) used the content based approach as shown in Figure 4. In

Table 2: List of events where articles were published.

#	Conference
1	ACM Conference on Recommender Systems (RecSys)
2	ACM Conference on User Modeling, Adaptation and Personalization (UMAP)
3	ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
4	ACM workshop on Research advances in large digital book repositories and complementary media
5	Brazilian Symposium on Multimedia and the Web (WebMedia)
6	China National Conference on Chinese Computational Linguistics International Symposium on Natural Language Processing Based on Naturally Annotated Big Data (CCL2015, NLP-NABD 2015)
7	IC3K 2013; KDIR 2013 and KMIS 2013
8	IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)
9	IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
10	International Conference of the IEEE Engineering in Medicine and Biology Society
11	International Conference on Collaboration Technologies and Systems (CTS)
12	International Conference on Computational Collective Intelligence (ICCCI)
13	International Conference on e-Business Engineering (ICEBE)
14	International Conference on Knowledge Discovery and Information Retrieval (KDIR)
15	International Conference on Tools with Artificial Intelligence
16	International Conference on Web Intelligence (WI)
17	International Florida Artificial Intelligence Research Society Conference (FLAIRS)
18	International World Wide Web Conference Committee (IW3C2)
19	Italian Research Conference on Digital Libraries (IRCDL)
20	Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)
21	Knowledge Engineering and Ontology Development conference (KEOD)
22	LWA joint conference
23	Symposium on Network Cloud Computing and Applications (NCCA)
24	Workshop on Bibliometric-enhanced Information Retrieval (BIR)

Table 3: List of journals where articles were published.

#	Journal
1	The Data Base for Advances in Information Systems
2	Decision Support Systems
3	Frontiers in Artificial Intelligence and Applications
4	IEEE Access
5	IEEE Transactions on Big Data
6	International Journal of Technology Enhanced Learning
7	Journal of Intelligent Information Systems
8	Journal of Systems and Software
9	Mobile Networks and Applications
10	Procedia Computer Science

addition to the traditional approaches, there were also hybrid studies with seven studies (18.4%), as well as a knowledge-based research (2.6%). Other different models (7.9%) includes: Markov Chain Based Model, Ontology Based Model, and Time Context Based Model.

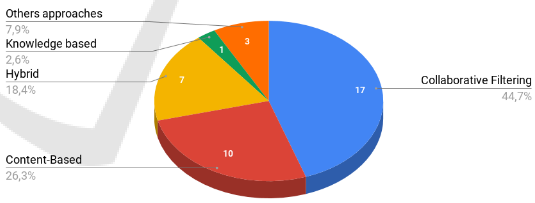

Figure 4: Main Approaches in Recommender System.

As can be seen in Figure 4, almost half of the approaches studied applied the Collaborative Filtering model. In order to investigate further, charts were made with the subcategories of this model, we have identified the occurrence of four model-based approaches, two of them were approaches that used Machine Learning algorithms and the other two approaches used Matrix Factorization. Finally, thirteen studies used the Neighborhood-based approach as seen in the subdivisions of approaches in the chart in Figure 5.

A traditional approach which was also widely used is the content-based approach with ten studies. They were also divided according to their sub-

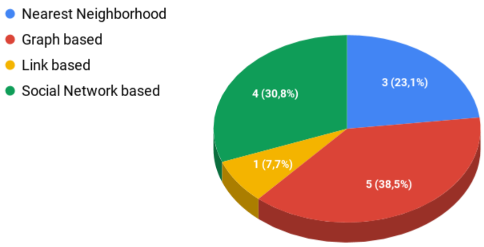

Figure 5: Neighborhood-based Approaches.

classification, with the presence of eight studies that used mathematical models and only two studies using Machine Learning algorithms.

In addition to studies with unique approaches, hybrid approaches were also cited resulting in seven studies as can be seen in Figure 6, highlighting the so-called 'mixed' approaches, which are based on the merging and presentation of multiple classification lists in only one, that is, the main algorithm will produce recommendation lists that can be merged into a single classified recommendation list Taghavi et al. (2017). Hybridization of recommendation systems combine techniques for higher performance, trying to use the advantages of one technique to correct the disadvantages of another.

For example, in collaborative filtering there is the problem called Cold-Start, in which the system is unable to recommend new unrated items Jannach et al. (2010). The content-based approach does not face this kind of problem, as recommendations are based on the content of the items that are most easily available Ricci et al. (2015). Presumably, most of the hybrids approaches merge the collaborative and content-based filtering techniques.

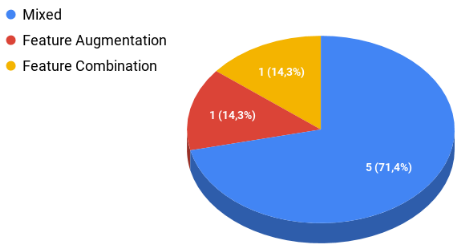

Figure 6: Hybrid Approaches.

Besides the mixed approaches, there are studies using Feature Augmentation and Feature Combination with one study each. The difference between these two approaches is that in Resource Combination there are two components: the actual recommending system, and the contributing system of the recommending system. The contributing component inserts resources into the recommending system source, and the recommending system work with data modified by the contributing system. The Feature Augmentation hybrid is similar to the Feature Combination hybrid, however, it is more flexible and adds smaller dimensions as the contributor produces new features.

3.5 Which Databases Were Used?

SRQ5 aims to identify which databases were used for the studies, as researchers needed a large amount of academic data. For this reason, most databases are based on scientific search engines, indexers, and digital libraries. The most used was the base *CiteULike* with 10 uses shown in Figure 7. There was also emphasis on the use of proprietary databases by the authors, being impossible to say the origin of the data that were used. The databases that contained only 1 use were grouped into Other, among them there are: *ACL Anthology Reference Corpus*, *ACM DL*, *ArXiv*, *BDBComp Digital Library*, *Dspace Publication Database*, *LitRec*, *Mendeley*, *Scopus*, *Scholarly Publication Recommendation Dataset*, among others.

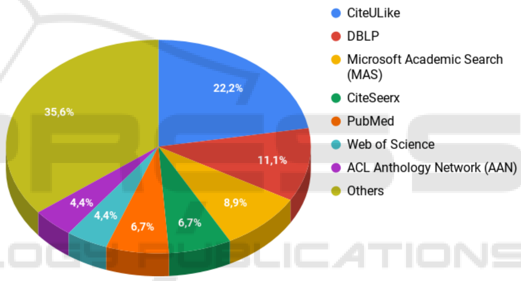

Figure 7: Most used databases.

3.6 Which Algorithm(s) are Being Used?

The methods and algorithms applied to generate, or even improve, recommendation systems are as diverse as possible. Therefore, to respond to QS6, the algorithms used by the authors were observed and categorized using the taxonomy of the recommendation systems development phase proposed by Taghavi et al. (2017). This taxonomy classified the methods and algorithms according to similar execution modes as can be seen in Figure 8, accounting for a total of 58 algorithms used in the studies. Often, the algorithms had a few variations between models because they were versions of the same algorithm model.

In content-based RS approaches, Machine Learning algorithms are used, as well as Vector-based Representation algorithms. In contrast, collaborative filtering-based RSs work primarily with Neighborhood Methods, representing the largest percentage

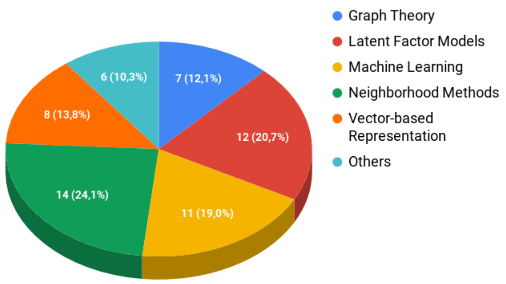

Figure 8: Most Used Algorithm Categories.

among studies: 24.1%. In Figure 9 we see the subcategories of neighborhood-based methods. The most used is Similarity Measure based algorithms with seven studies divided into Rating-based or Ranking-oriented algorithms. Classification-based algorithms used the following metrics: Cosine Similarity, Google Distance Similarity, and Pearson Algorithm.

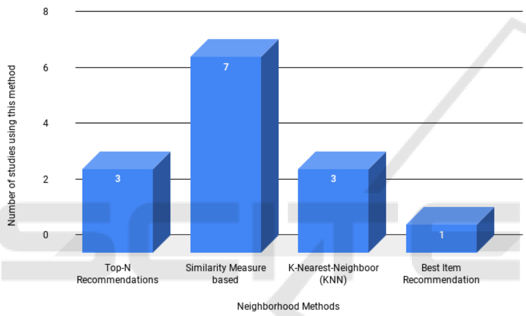

Figure 9: Neighborhood based methods.

Among the other methods, Top-N Recommendations used the Randomwalk algorithm, K-Nearest-Neighboor (KNN) used the algorithm of the same name, and Best Item Recommendation used the neighbor-weighted algorithm.

In addition to neighborhood-based methods that are typically Memory-based approaches, Model-based approaches are represented by Latent Factor Model methods which can be verified as shown in Figure 10. These include probabilistic models that apply Machine Learning algorithms, besides an article that used the KNN algorithm. Matrix Factorization methods employed different types of algorithms such as: Exposure Matrix Factorization, Singular Value Decomposition, and Time-aware Factor Model.

In Figure 11, it is possible to verify Machine Learning methods based on the use of classification algorithms: Naive-Bayes Classifier, Support-Vector Machine Classifier (SVM), and the proprietary algorithm *Cavnar-Trenkle*, as well as the use of clustering algorithms: K-Means and K-Medoids.

Other methods also identified were a study based on the MapReduce process, studies based on the

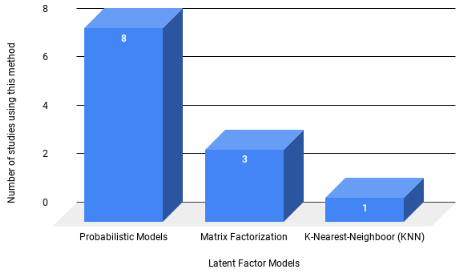

Figure 10: Latent Factors Methods.

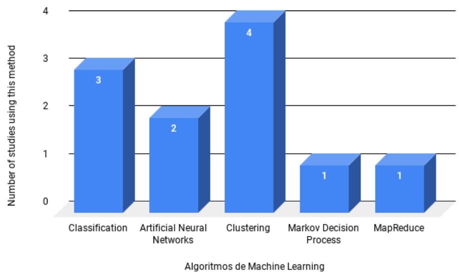

Figure 11: Machine Learning Algorithms.

Markov Decision Process and two studies on the Artificial Neural Networks (ANN) using Recurrent Neural Networks (RNN).

As mentioned here, in addition to Machine Learning algorithms, as content approaches as well, we use vector representation using methods such as the TF-IDF (Measure Frequency Inverse Measure Frequency) metric and also algorithms such as the *Rocchio* Algorithm with 6 and 2 models respectively.

The rest of the Recommender System approaches use mostly the same algorithms as the collaborative methods and the content-based methods cited. In addition to that, they also have the methods derived from graph theory as can be seen in Figure 12. Among the algorithms found in the studies were: Page Rank Algorithm, Community Partition Algorithm, Greedy-order Algorithms, HITS Propagation Algorithm, and the Graph-based Ranking.

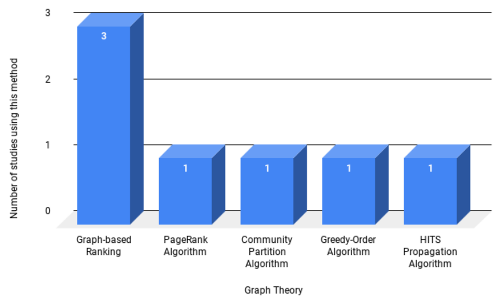

Figure 12: Graph theory based methods.

3.7 How is the Recommendation Process Evaluated in Terms of Metrics?

Seeking to respond to QS7, which aims to know the ways of evaluating a recommendation system, we sought to verify through studies which metrics were used to measure accuracy, recovery time, acceptance rate, and etc. In the investigated studies, it was found that three of the thirty-eight studies had not yet evaluated or did not mention how the recommendation system was evaluated. Evaluation metrics are critical to verifying that the approach used is working well and how we can improve the system. Figure 13 shows the types of metrics most used by the studies. Single-use metrics were grouped into Other, such as Confidence, Diversity, Novelty, Robustness / Stability, Scalability, Serendipity, and Confidence. Although similar terms appear as Confidence and Trust, they have different meanings. The Confidence metric is defined as the reliability of the recommendation and the system's confidence in its recommendations and / or predictions, and can be reported by the system confidence score Taghavi et al. (2017). The Trust metric refers to users' confidence in the recommendations provided by the system.

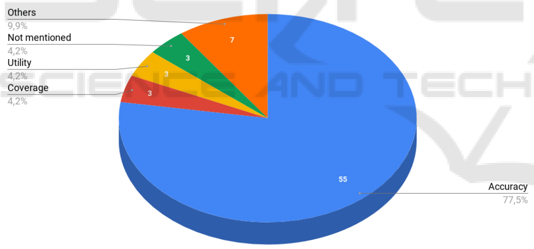

Figure 13: Metrics used in the studies.

In addition to the less used metrics previously mentioned with each use, there was also the application of metrics such as Coverage using algorithms such as Measure User Coverage (UCOV) and also metrics to measure the Utility presented by the recommendation system. However most studies have applied the use of metrics to measure the system's Accuracy which is divided into three subcategories: Ranking measures/Ranking accuracy accuracy), Relevance measure/Classification accuracy, and Rating Prediction Accuracy. In Tables 4 and 5 it is observed which measures were used in terms of the types of accuracy metrics.

The last subcategory Rating Prediction Accuracy used the Mean Percentile Rank (MPR) and Root Mean Square Error (RMSE) metrics.

Table 4: Accuracy Metrics- Ranking Measures/Rank Accuracy.

Ranking Measures/Rank Accuracy
Average Reciprocal Hit Rate
Normalized Discounted Cumulative Gain (NDCG)
Mean Reciprocal Rank (MRR)
Normalized Distance-based Performance Measure

Table 5: Accuracy Metrics - Relevance Measure/Classification Accuracy.

Relevance Measure/Classification Accuracy
F1
Mean Average Precision (MAP)
Mean Average Weighted Precision (MAWP)
Precision
Recall
Recall x Precision (interpolated)

4 CONCLUSION

The study of the Recommender Systems area through a systematic literature mapping allowed us to identify the state-of-the-art of Recommender Systems focusing on scientific publications. It is clear that the area is relatively new given the number of articles growing in recent years, and with a half-life[2] of 10 years. In addition, there several recommender systems with similar purposes and classic approaches, however, always seeking to improve these systems with new methods and algorithms. Among the new approaches that have been used is the recurrent neural network model, with good results, but only with further work will be able to prove the efficiency of the model. Among the most used databases were CiteULike and DBLP, although DBLP is not a database but an indexer of other databases.

The use of Natural Language Processing (NLP) in Recommender Systems in studies is still small, proving to be a good research gap, confirming the expectation that it is a prominent area. Thus, as a future work we intend to develop a Recommender System based on scientific search engines (SSEs), recommending publication vehicles for researchers, so that they know where to publish their scientific works based on the textual data of their articles.

[2]Estimated elapsed time for an article to receive half of all citaations it will have throughout its lifetime Diniz (2013).

REFERENCES

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, 17(6):734–749.

Brunialti, L. F., Freire, V., Peres, S., and Lima, C. A. d. M. (2015). aprendizado de máquina em sistemas de recomendação baseados em conteudo textual uma revisao sistematica. In *XI Brazilian Symposium on Information System*, pages 203–210, Goiânia. Association of Information Systems.

Buchinger, D., de Siqueira Cavalcanti, G. A., and da Silva Hounsell, M. (2014). Mecanismos de busca acadêmica: uma análise quantitativa. *Revista Brasileira de Computação Aplicada*, 6(1):108–120.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370.

Burke, R. (2007). Hybrid web recommender systems. In *The adaptive web*, pages 377–408. Springer, Switzerland.

Champiri, Z. D., Shahamiri, S. R., and Salim, S. S. B. (2015). A systematic review of scholar context-aware recommender systems. *Expert Systems with Applications*, 42(3):1743–1758.

Diniz, E. (2013). Editorial. *Revista de Administração de Empresas*, 53:223 – 223.

Ibrahima, O. A. S. and Younisb, E. M. (2018). Recommender systems and their fairness for user preferences: A literature study.

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press, UK.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.

Malhotra, R. (2015). A systematic review of machine learning techniques for software fault prediction. *Applied Soft Computing*, 27:504–518.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16):7653–7670.

Palaniappan, R., Sundaraj, K., and Ahamed, N. U. (2013). Machine learning in lung sound analysis: a systematic review. *Biocybernetics and Biomedical Engineering*, 33(3):129–135.

Park, D. H., Kim, H. K., Choi, I. Y., and Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11):10059–10072.

Petersen, K., Vakkalanka, S., and Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18.

Pons, E., Braun, L. M., Hunink, M. M., and Kors, J. A. (2016). Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343.

Portugal, I., Alencar, P., and Cowan, D. (2017). The use of machine learning algorithms in recommender systems: a systematic review. *Expert Systems with Applications*.

Ricci, F., Shapira, B., and Rokach, L. (2015). *Recommender systems handbook, Second edition*. Springer, Switzerland.

Skiena, S. S. (2017). *The Data Science Design Manual*. Springer, Switzerland.

Taghavi, M., Bentahar, J., Bakhtiyari, K., and Hanachi, C. (2017). New insights towards developing recommender systems. *The Computer Journal*, 61(3):319–348.

Tan, P.-N., Steinbach, M., and Kumar, V. (2009). *Introdução ao Data Mining Mineração de Dados*. Ciência Moderna, Rio de Janeiro.

Wazlawick, R. (2017). *Metodologia de pesquisa para ciência da computação*, volume 2. Elsevier Brasil, Rio de Janeiro.