A Semantically Enriched Recommendation and Visualization Approach for Academic Literature

Corinna Breitinger
Dept. of Computer and Information Science
University of Konstanz
Konstanz, Germany
corinna.breitinger@uni-konstanz.de

ABSTRACT

Researchers must be aware of the current developments and novel findings in their field. To gain an overview of relevant prior work and ongoing research, academics must review the published literature. However, the search for related academic literature is tedious. Furthermore, researchers can easily oversee potentially valuable information in today's increasing volume of academic literature. While academic search and recommendation engines have greatly simplified the information acquisition process, current recommendation approaches are not adequately considering specialized semantic similarity measures, such as citation-based similarity, mathematical formulae-based similarity, or image-based similarity to recommend and visualize academic literature. This paper proposes to take into account combinations of semantic features that have previously not been considered for the use case of literature recommendation.

Additionally, supporting researchers in the sense-making of semantic similarities present in recommended literature remains a largely unsupported task. Researchers must review the content of each academic paper, manually identify or compare the sections that are of interest to them, and then arrive at a judgment regarding the relevance of the recommendation. We propose a supportive process that allows researchers to more quickly compare academic literature with regard to the user-selected semantic features that are of interest to them. Such user-selected features can be the citations to other literature, the contained mathematical formulae, graphs, or figures, i.e. a range of semantic features that go beyond pure text-based similarity. The proposed *semantically-enriched* literature recommendation and visualization concept could help researchers more quickly identify the specific content that is of interest to them within a large set of recommended literature.

Keywords

Information retrieval, recommender systems; semantic analysis; information visualization

ACM Reference format:
C. Breitinger. A Semantically Enriched Recommendation and Visualization Approach for Academic Literature. In *Bulletin of the IEEE Technical Committee on Digital Libraries (TCDL)*, vol. 15, iss. 1, 2019.

1. INTRODUCTION

Approximately 2.5 million new scientific papers are published each year [40]. At the same time, the number of scientific publications has been increasing at a rate of 8-9% annually, which translates to a doubling of global scientific literature approximately every nine years [6]. With so much information available, *literature recommender systems* (LRS) are a crucial filtering and discovery tool used by academics to identify the most relevant literature.

LRS can support researchers by pointing them to related work in their field, which they may have otherwise missed. For example, effective LRS can prevent researchers from performing duplicate or suboptimal research because they were unaware of research that has already been published in their field. Instances of redundant research remain a regular occurrence in academia [24]. Far more damaging than redundant research, however, is when researchers rely on outdated findings, or findings that are proven false by a more recent publication. Especially in the medical field, such misinformation can be harmful. These problems could potentially be prevented if researchers were able to more easily identify and more efficiently browse the academic literature that is relevant to their information needs.

Citation-based approaches and text-based document similarity measures are commonly being employed by today's LRS [23,41]. However, today's LRS share two shortcomings.

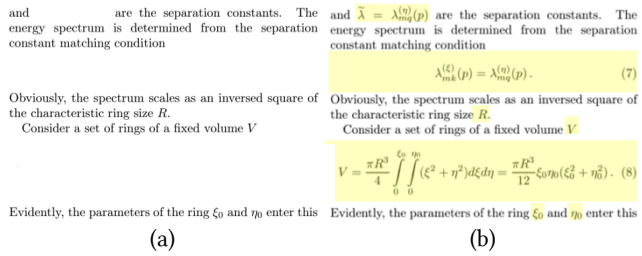
(1) They do not sufficiently consider the semantic features that are unique to academic literature in order to determine relevance, i.e. they are not adequately considering semantic analysis approaches, such as the analysis of in-text citation placement, the similarity of mathematical language, or the similarity of figures and images.

(2) Today's LRS provide no visualization concept for the types of semantic features contained in academic literature. Semantic similarity is typically more subtle than text-based similarity (i.e. verbatim text matches) and is thus especially time-consuming and challenging for users to identify upon a quick reading of the text. However, if adequately highlighted, users could be supported in better understanding the semantic relevance present in the recommended research papers. For example, a visualization of

semantic features could quickly show the user *where* in an academic publication the similar features occur, either with regard to a user query or with regard to a reference publication as input.

When it comes to improving both literature recommendation generation and sense-making by the user, current LRS are largely ignoring recent advancements in specialized semantic similarity analysis. However, especially in the STEM fields (Science, Technology, Engineering and Mathematics), taking non-text-based semantic features, such as formulae, citation patterns, graphs, or figures into consideration can be extremely valuable for a more complete understanding of an academic article's content.

Figure 1 shows two variations of the same text section extracted from a STEM publication. Excerpt (a) on the left only shows the plain text, without mathematical expressions, while excerpt (b) on the right shows the same text with the equations included as they were intended by the researchers. It becomes clear that in the absence of the mathematical expressions, a significant chunk of the semantic information goes missing.


and are the separation constants. The energy spectrum is determined from the separation constant matching condition

Obviously, the spectrum scales as an inversed square of the characteristic ring size R.
Consider a set of rings of a fixed volume V

Evidently, the parameters of the ring ξ_0 and η_0 enter this

(a)

and $\widetilde{\lambda} = \lambda^{(\eta)}_{nnq}(p)$ are the separation constants. The energy spectrum is determined from the separation constant matching condition

$$\lambda^{(\xi)}_{mk}(p) = \lambda^{(\eta)}_{nnq}(p)\,. \quad (7)$$

Obviously, the spectrum scales as an inversed square of the characteristic ring size R.
Consider a set of rings of a fixed volume V

$$V = \frac{\pi R^3}{4}\int_0^{\xi_0}\int_0^{\eta_0}(\xi^2+\eta^2)d\xi d\eta = \frac{\pi R^3}{12}\xi_0\eta_0(\xi_0^2+\eta_0^2)\,. \quad (8)$$

Evidently, the parameters of the ring ξ_0 and η_0 enter this

(b)

Figure 1: Current LRS ignore text-independent semantic markers, such as mathematical expressions (a). However, such expressions are an essential component for the semantic understanding of many publications, especially in the STEM fields (b).

Figure 1 thus illustrates the importance of also taking into account non-textual semantic content, in this case, mathematical formulae. By ignoring such non-textual semantic features from the analysis, existing approaches lose out on significant information-bearing features that encode precise and semantically-rich content. We hypothesize that by taking into consideration semantic features that go beyond text-based similarity, the recommendation quality of LRS can be significantly improved – in particular for academic literature in the STEM fields.

2. RESEARCH OBJECTIVES

In contrast to a text's literal similarity, semantic similarity is the likeness of meaning in the absence of literal text matches. Common semantic features present in academic literature include citation patterns, images, graphs, figures, chemical formulas, as well as mathematical expressions.

While some of these features have been analyzed individually for their suitability as similarity measures, existing academic literature recommendation systems are not fully taking these semantic features into account.

2.1 Objective

The shortcomings of academic literature recommendation systems, motivated me to define the following research objective:

Conceive, implement, and evaluate a recommendation approach that takes into consideration text-independent semantic features, such as academic citations, mathematical formulae, and figures, to generate semantically-enriched recommendations tailored to academic literature in the STEM fields.

The goal of the planned research is twofold. First, to conceive and prototype a semantically-enriched literature recommender that augments today's content-based recommendation approaches with more granular semantic similarity approaches to account for the unique instances of semantic feature similarity that exists in STEM literature.

Second, a visualization concept must be devised, which for the first time enables users to more quickly make sense of the semantically similar features contained in a set of recommended literature.

We hypothesize that by taking into consideration a document's semantic characteristics, users can be better supported in the discovery of semantically relevant literature beyond text-based similarity alone. For example, recommending academic literature that contains similar features, or similar patterns of features, such as citations, formulae, and figures could make apparent these more subtle semantic similarities, which current literature recommender systems are not sufficiently taking into account.

2. 2 Research Tasks and Questions

To address this research objective, we have defined the following research tasks.

1.) Review today's literature recommender approaches with a special focus on answering:

   a. What is done to improve the recommendation quality for the less text-heavy literature from the STEM fields?

   b. What characteristics beyond textual similarity are being used to improve recommendations?

   c. What are the special requirements for literature recommendation in the STEM field?

2.) Conceive and design a recommendation approach that considers the features unique to publications in the STEM fields by giving special consideration to non-textual characteristics such as formulas, figures, and citations. Adapt existing suitable approaches from the research domain of plagiarism detection.

3.) Implement the novel approach in a literature recommender system.

4.) Evaluate the recommender system in regard to its performance (recommendation quality, i.e., addressing the information need) and its usability. Semantic similarities

might not be obvious and need to be communicated to the user effectively and efficiently.

5.) Give recommendations on an ideal recommender system by considering the outcomes of the evaluations and user study. Derive appropriate weighting for the different similarity measures depending on the user (his or her information need), and the research field (e.g., LRS for math-heavy publications should place a higher weight on formulas and a lower weight on text-based similarity).

Naturally, the most representative semantic features and their frequencies of occurrence will differ among the STEM disciplines. For example, assessing the semantic equivalence of mathematical formulae will likely be valuable for mathematics or physics literature. However, determining the presence of citation-based similarity will likely be applicable to all fields. The most typical semantic features and their frequencies and combinations for the different STEM fields is still very much an open research question to be addressed (cf. Research Task 2).

The final step of visualizing the identified forms of semantic similarity within the recommended literature will be especially valuable since instances of semantic similarity are often more subtle than text-based similarity. This makes any similarities among semantic features more difficult for users to manually identify upon first glance. The proposed research will result in the design and development of a visualization concept to enable users to quickly navigate and make sense of the identified semantic similarities within a set of recommended literature. With the help of user studies, I plan to evaluate the recommendation effectiveness and the devised semantic visualization concept for the proposed *Semantically-enhanced Literature Recommender*.

3. STATE OF THE ART

3.1 Academic literature recommendation

Existing recommender systems for academic literature employ either *content-based* filtering (CBF) approaches, *user-based* approaches (e.g., *collaborative filtering* (CF)), *graph-based* approaches, or a combination of these three types of recommendation approaches [4]. In a review of 62 approaches for research paper recommendation, we found that the majority of reviewed systems (55%) used *content-based* approaches to recommend related academic literature [4].

The quality of content-based recommendations by design depends on the level of comprehension of the recommended item's content. Thus, taking into account not only a document's textual similarity and citation graph but also all *semantic features*, as well as the placement of these features, is a crucial consideration for further improving CBF recommendation approaches.

User-based approaches alone, such as CF, were used only by a minority of academic literature recommender systems (18%) [4]. When applied to the academic publication recommendation use case, CF faces a set of significant drawbacks, such as the cold-start problem [37], since CF requires high levels of user participation, but readers' motivation to consistently interact within an LRS is low. This circumstance makes CF a rather unsuitable approach for recommending academic literature since any implicit or explicit ratings that can be collected are too sparse. In comparison to the product recommendation or movie recommendation use cases, for the academic literature recommendation use case there exist millions of papers, but very few readers.

3.2 Semantic similarity measures

A wide range of semantic analysis methods have been proposed to determine document similarity. However, many state-of-the-art semantic similarity assessment methods have thus far only been applied to a single specialized use case. Figure 2 gives an overview of semantic similarity measures and the features they consider. In the category, which we term *semantic marker-based approaches* for similarity determination, *citation-based approaches* have been used to detect heavily disguised plagiarism [16,17,26]. *Formula-based methods* have been used to determine the semantic equivalence of mathematical formulae for Mathematical Information Retrieval (MathIR) tasks and also to detect mathematical plagiarism (MathPD) [28]. *Named Entity (NE)* analysis considers patterns of semantic markers within the full text, similarly to citation-based methods, and has recently been proposed for uncovering media bias in news articles [19]. Finally, *image similarity measures* are used for image classification and

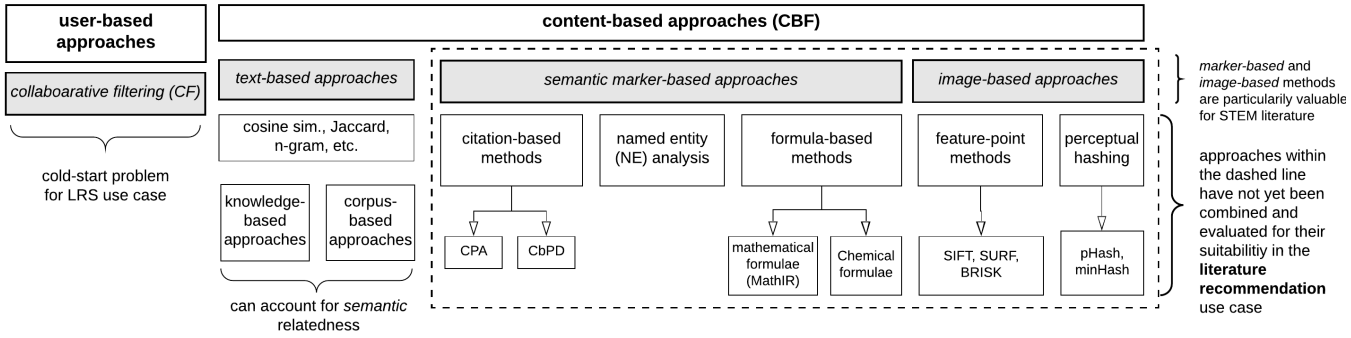


Figure 2: Overview of semantic document analysis approaches. Methods in the dashed outline have thus far not been combined and examined for their suitability in the LRS use case.

near duplicate image detection tasks, including detecting plagiarism [8,27] or battling image spam [13].

In the following, I review the state-of-the-art approaches that appear most promising for improving the semantic understanding of academic literature in the STEM fields. I have classified semantic document analysis methods into the categories: text-based approaches, marker-based approaches (which are natural language independent), and image-based approaches. Note that I will not review well-known text retrieval methods, such as cosine similarity, Jaccard index, n-gram string matching, etc. since these rely on literal word matches. Our aim is to determine the presence of semantic similarity independent of literal text-based similarity.

Text-based approaches capable of accounting for *semantic relatedness* commonly use knowledge-based or corpus-based approaches [30], or a combination of both [10]. Knowledge-based approaches make use of encyclopedias or thesauri that specify the linguistic relations between words, such as WordNet[1], VerbNet[2], PropBank[3], or FrameNet[4]. These lexical databases are commonly used for query expansion [39,42]. They have also been used in the field of plagiarism detection to enable a semantic matching of text beyond literal keyword-based matching [7].

Corpus-based approaches, on the other hand, assume that semantically related words occur in similar contexts and make use of very large corpora to automatically extract the semantic relations. Latent Semantic Analysis (LSA) [10] is one such corpus-based approach, which learns semantic relations from word co-occurrence patterns within a corpus. A weakness of LSA is that it depends heavily on the characteristics of the corpus and does not make use of human expert knowledge. An approach, which addresses this weakness of LSA, is Explicit Semantic Analysis (ESA) [12]. By using, for example, the Wikipedia corpus as a semantic background, ESA can be used on a range of academic topics. Recently, ESA has been applied to the plagiarism detection use case [29]. Lastly, cross-language approaches (e.g., using machine translation, or knowledge graphs [11]) have been proposed for semantic similarity analysis.

Citation-based approaches, such as Co-Citation Proximity Analysis (CPA) [3] or Citation-based Plagiarism Detection (CbPD) [16], exploit the citation patterns present in academic texts as language-independent features to identify semantic similarity even in the absence of literal text matches. The CbPD approach has been successfully applied to identifying heavily modified plagiarism in real-world, large-scale collections [16,25]. A hybrid approach was subsequently developed, which combines the CbPD method with traditional character-based (i.e., text-based) heuristics to retrieve candidate documents from a reference collection. The hybrid approach sharply reduced computational time in the plagiarism detection use case, since citation-based methods are far more efficient on very large document collections than the computationally expensive text-string comparison methods [25].

Mathematical Information Retrieval (MathIR) is a specialized subdomain of document semantic analysis that uses Mathematical Language Processing (MLP) to assess the semantic relatedness of documents [32,34,35]. One of my colleagues, Moritz Schubotz, pioneered effective approaches for MLP in his Ph.D. thesis [34]. For this highly specialized technical domain, I will make use of the prior work performed by my research group members Moritz Schubotz and Norman Meuschke. I also plan to collaborate with Philipp Scharpf, who is researching how to improve formula-based methods for a better semantic understanding of STEM literature.

Image-based similarity analysis approaches are varied [38], with different approaches being more or less suitable for identifying similar features either in photographs, abstract figures, or graphs. Well established feature-point methods include SIFT [22], SURF [1], and BRISK [21], which are capable of retrieving slightly discrepant versions of the same image. Perceptual hashing (pHash) [18], or min-Hash [8] represent another class of approaches in which images perceived as similar by humans result in similar hash values for image similarity determination. My colleagues and I recently applied perceptual hashing in an adaptive image-based plagiarism detection approach [27]. However, to the best of my knowledge, image-based approaches have thus far not been considered for enhancing academic literature recommendation.

In summary, marker-based semantic similarity analysis methods have been applied to several use cases, including the plagiarism detection (PD) use case. However, these approaches have not been examined in combination in order to provide a semantically enriched recommendation concept tailored to the characteristics of STEM literature. My aim is to determine a suitable combination and weighting of existing *marker-based* and *image-based* assessment approaches and apply them to the LRS use case. As mentioned previously, STEM literature can contain fewer literal text-based similarity than in other fields, while exhibiting higher semantic marker-based similarity. For example, a STEM publication may contain similar citation patterns or mathematically equivalent expressions, as well as image-based similarity, e.g. similar charts, figures, or image data generated by lab equipment.

3.3 Visualization of Recommended Literature

Visualizations to support the discovery of semantic similarities in documents already exist for the PD use case [26]. However, to the best of my knowledge, no *semantically-enriched visualization concept* exists for the literature recommendation use case.

Figure 3 shows the typical interfaces of today's LRS. Readers receive no visualization of the contained semantic content beyond

[1] https://wordnet.princeton.edu/
[2] https://verbs.colorado.edu/~mpalmer/projects/verbnet.html
[3] https://verbs.colorado.edu/~mpalmer/projects/ace.html
[4] https://framenet.icsi.berkeley.edu/fndrupal/

recommendation lists, which commonly only display a publication's metadata and citation information.

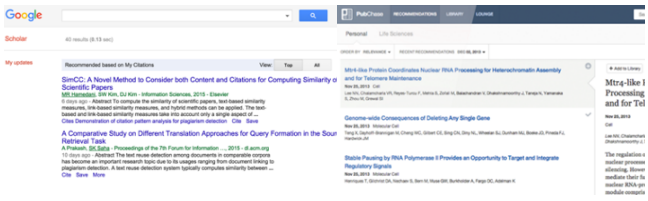


Figure 3: Current LRS interfaces offer no similarity visualization of a document's content (left: Google Scholar, Right: PubChase)

Today's leading academic recommenders, including the personalized recommendations of Google Scholar, Mendeley Suggest, Docear, PubChase, CiteULike, and the 'Similar Papers' section in Semantic Scholar, provide list-based interfaces.

Mutlu et al. argue that in cases when users must compare and relate information, recommendation lists become incomprehensible and tedious [31]. Collins et al. discuss the inherent "position bias" of list-based literature recommendations, which describes the tendency of users to interact with items at the top of a list with higher probability than with items at lower positions, regardless of the items' actual relevance [9].

A more detailed visualization concept capable of pointing out the semantic content of publications might help users to more quickly arrive at a judgment regarding the relevance of recommendations. To the best of my knowledge, no such semantically-enhanced visualization concept has thus far been proposed to support users in making sense of the semantic content present in recommended academic literature.

In contrast to list-based exploration, Figure 4 shows the UI of the PD system, CitePlag [16]. CitePlag implements a citation-based approach (CbPD) to identify matching citation patterns in academic literature as a potential indicator for plagiarized content.

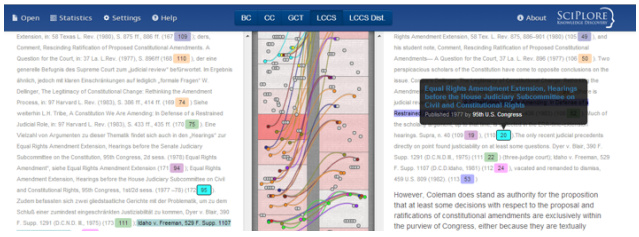


Figure 4: Visualization of citation-based similarity in the plagiarism detection system CitePlag (http://citeplag.org/)

In Figure 4, matching in-text citations are marked and connected in an interactive pattern visualization column (center). By clicking on the connected dots, users can quickly navigate to sections with high citation pattern similarity, i.e. in this case, sections which are potentially plagiarized. By considering textually independent citation patterns, even heavily paraphrased and translated text sections remain identifiable if the plagiarism shares suspicious overlap in its citation patterns with its original source [15].

The visualization concept in Figure 4 shows how instances of matching semantic features (in this case, academic citations) are harder for humans to identify when compared to literal text similarity. For the LRS use case, this means a suitable visualization concept may better support users in identifying the more subtle semantic relevance contained in the recommended literature. Before users begin reading recommended documents in their entirety, they could benefit from a visualization concept similar to the one I helped conceive for the PD use case in CitePlag system [16,26].

Sinha and Swearingen stress the importance of approaching recommender systems from a user's perspective and taking into consideration user interface issues [36]. A visualization concept would be especially crucial for the quick identification of citation-based, formula-based, or figure-based semantic similarity that is common to STEM literature. Despite advances in user-adaptive recommendation interfaces [20], and recommenders explaining the reason behind recommendations [14,36], the interfaces of LRS are lacking when it comes to making sense of the semantic relevance of the recommended literature.

Conceiving a visualization concept that will enable users to identify and browse the identified marker-based and image-based instances of semantic similarity will thus be the secondary aim of the presented research.

4. TOWARD SEMANTICALLY-AWARE LITERATURE RECOMMENDATIONS

4.1 Semantically-enriched Recommendation Generation

In the first stage of research, I will review, combine, and adapt existing approaches that consider the features unique to publications in the STEM fields. As stated previously, the focus lies on non-textual characteristics, such as mathematical formulae, figures, and citation patterns. Many of these approaches have already shown promise in the plagiarism detection use case, a field of research with which I am already familiar. Conceiving and designing a recommendation approach that most effectively combines existing semantic similarity measures and applies them in a content-based recommendation approach for academic literature will be the outcome of this first research stage (Task 1).

The novel recommendation approach will then be implemented in an LRS prototype (Task 2). I have already begun working on the acquisition of a suitable recommendation corpus using the large-scale real-world academic literature datasets PMC OAS and arXiv. PMC OAS[5] will supply literature from the Life Sciences and Biomedical field. For the purpose of corpus creation, I am adhering

[5] ncbi.nlm.nih.gov/pmc

to the broader definition of STEM subjects by the NSF, which also includes the Life Sciences. ArXiv[6] will supply literature from Physics, Mathematics, Statistics, Quantitative Biology, Computer Science, and Electrical Engineering. This choice ensures a broad spectrum of STEM fields.

Depending on the research field and the user's information need, the most suitable combination and weighting of text-independent semantic similarity features will need to be determined. For example, molecular biology and other lab sciences may require image-based measures to play a larger role, while physics will likely require formulae-based measures to play a larger role. Additionally, the order in which mathematical formulae and image-based similarities occur within the full-text of a document should also be taken into consideration, analogous to the CbPD approach for citation pattern analysis.

Methods of cross-language analysis have been excluded as a result of an initial review of methods. However, they may be examined for future integration into the LRS, if the system is expanded with suitable cross-language corpora.

4.2 Semantic Visualization Concept

Semantic similarities are typically less obvious to readers and must thus be communicated effectively. The second research aim is thus developing and evaluating a visualization concept for the marker-based and image-based relevance present in the recommended literature (Task 3).

To achieve this, a web-based user interface will be developed and evaluated with the help of user studies. Based on the feedback of participants, I will further improve upon the visualization concept. An initial system requirements elicitation has already taken place. Such an interface and visualization concept, when applied to the recommendation use case, must show to its users which forms of semantic similarities are present. It must also allow users to quickly navigate to the instances where these similarities occur within an article's full text. One of the paper-based design prototypes shown to participating researchers in the initial requirements elicitation study is shown in Figure 5.

The LRS will include a side-by-side visual comparison of two or more document full-texts or individual chapters. Users will have the ability to filter for individual semantic features of interest and to quickly scroll through the paper by clicking on the highlighted semantic features. Visualization components, such as those presented by Reiterer et al. in the INSYDER prototype [33], may also be integrated to support the user in the visual semantic-similarity-browsing task.

Furthermore, the literature recommendation browsing experience will be customizable by users depending on their academic discipline and their changing information retrieval needs. The recommender system will be evaluated with the help of a user study to assess recommendation quality, i.e. its ability to address a user's specified information need, and its overall usability (Task 4).

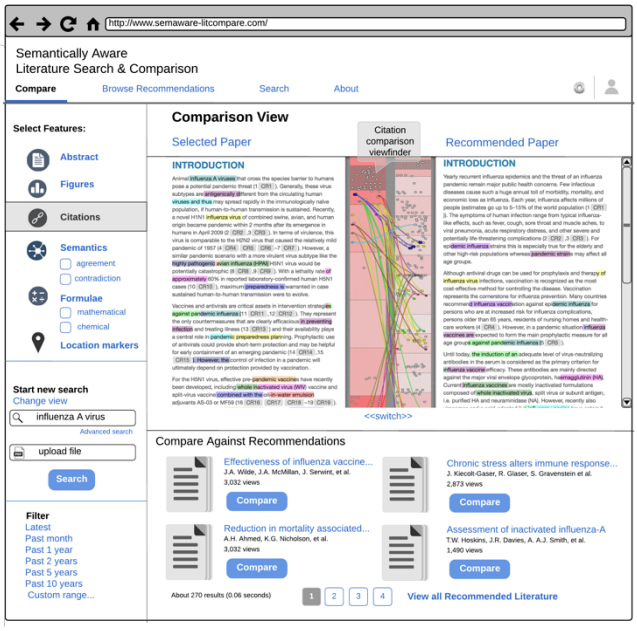


Figure 5: Example of a semantically-enhanced visualization concept for browsing recommend STEM literature. Here, the user has selected the tab 'citations' as the semantic feature of interest.

5. PRELIMINARY WORK

The time required to familiarize myself with the state-of-the-art and current approaches in this field has been significantly reduced thanks to my prior work on citation-based approaches and academic recommender systems. I have already researched semantic similarity detection methods, in particular methods for citation-based similarity applied to the PD use case [16], and methods for image-based plagiarism detection [27].

In the field of recommender systems, I worked on the Docear[7] mind mapping project, where I evaluated user modeling approaches for extracting the content and characteristics of mind maps to generate literature recommendations [5]. In collaboration with Prof. Joeran Beel, I co-authored an extensive survey on academic research paper recommender systems [4], and I have researched and defined some of the challenges pertaining to academic recommender system evaluation in earlier work [2].

6. EXPECTED CONTRIBUTIONS

The presented research will make two key contributions. First, I will research and conceive a content-based literature recommendation approach capable of taking into account a document's semantic marker-based and figure-based similarities. Given the insights from the evaluation of this proposed

[6] arxiv.org

[7] http://www.docear.org/

semantically-enriched LRS, I will be able to provide recommendations on an ideal recommender system that is specifically tailored to the STEM fields. Second, I will conceive and evaluate a suitable visualization concept to support users in making sense of the semantic similarities contained in the recommended literature.

By improving semantic recommendation performance and by providing a semantically-enhanced visual literature comparison interface, researchers will be better supported in the literature discovery process. Researchers may also be able to better identify documents that fulfill highly specific and narrow information needs, which existing literature recommendation approaches cannot address.

The developed system will be provided as open source to allow extensions by the scientific community, e.g. by including additional semantic similarity measures, expanding the document collection, or customizing the system to specialized IR tasks.

7. CONCLUSION

In summary, academic literature recommendation systems could significantly benefit from taking into account a combination of marker-based and image-based semantic similarity measures. Especially in the STEM fields, considering non-text-based semantic features is vital to a complete understanding of a text. Determining a suitable combination and weighting of existing semantic approaches depending on the academic field and a user's information needs are at the heart of this research. The result of the proposed research will lead to the first semantically-enriched content-based recommendation approach tailored specifically to the characteristics of STEM literature.

Additionally, a visualization concept will be developed to provide readers of STEM literature with a more semantically-inclusive understanding of the recommended literature. The UI design and semantic similarity visualization concept will allow users to quickly identify, browse, and make sense of the different forms of semantic document similarity that occur within a set of recommended scientific literature.

ACKNOWLEDGMENTS

I thank the Carl-Zeiss Foundation for awarding me with a research scholarship that is funding my Ph.D. research.

REFERENCES

[1] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. Speeded-up robust features (SURF). *Computer vision and image understanding 110*, 3 (2008), 346–359.

[2] Beel, J., Breitinger, C., Langer, S., Lommatzsch, A., and Gipp, B. Exploring Reproducibility in Recommender-Systems Research. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research (UMUAI)*, (2016).

[3] Beel, J. and Gipp, B. Co-citation Proximity Analysis (CPA) – A New Approach for Identifying Related Work Based on Co-Citation Analysis. *12th International Conference on Scientometrics and Informetrics (ISSI'09)*, (2009).

[4] Beel, J., Gipp, B., Langer, S., and Breitinger, C. Research-paper Recommender Systems: A Literature Survey. *International Journal on Digital Libraries 17*, 4 (2016), 305–338.

[5] Beel, J., Langer, S., Kapitsaki, G., Breitinger, C., and Gipp, B. Exploring the potential of user modeling based on mind maps. In *User Modeling, Adaptation and Personalization*. Springer, 2015, 3–17.

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

[7] Chen, C.-Y., Yeh, J.-Y., and Ke, H.-R. Plagiarism detection using ROUGE and WordNet. *arXiv preprint arXiv:1003.4065*, (2010).

[8] Chum, O., Philbin, J., and Zisserman, A. Near Duplicate Image Detection: min-Hash and tf-idf Weighting. *BMVC*, (2008).

[9] Collins, A., Tkaczyk, D., Aizawa, A., and Beel, J. Position Bias in Recommender Systems for Digital Libraries. *International Conference on Information*, (2018), 335–344.

[10] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American society for information science 41*, 6 (1990), 391.

[11] Franco-Salvador, M., Gupta, P., and Rosso, P. Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing. In *Bridging Between Information Retrieval and Databases*. Springer, 2014, 227–236.

[12] Gabrilovich, E. and Markovitch, S. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *IJcAI*, (2007), 1606–1611.

[13] Gao, Y., Yang, M., Zhao, X., et al. Image spam hunter. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, (2008), 1765–1768.

[14] Gedikli, F., Jannach, D., and Ge, M. How should I explain? A comparison of different explanation types for recommender systems. *Int. J. Hum.-Comput. Stud. 72*, (2014), 367–382.

[15] Gipp, B., Meuschke, N., and Beel, J. Comparative evaluation of text-and citation-based plagiarism detection approaches using GuttenPlag. *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, (2011), 255–258.

[16] Gipp, B., Meuschke, N., and Breitinger, C. Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus. *Journal of the American Society for Information Science and Technology (JASIST) 65*, 8 (2014), 1527–1540.

[17] Gipp, B., Meuschke, N., Breitinger, C., Lipinski, M., and Nürnberger, A. Demonstration of citation pattern analysis for plagiarism detection. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, ACM Press (2013), 1119.

[18] Hadmi, A., Puech, W., Said, B.A.E., and Ouahman, A.A. Perceptual image hashing. In *Watermarking-Volume 2*. InTech, 2012.

[19] Hamborg, F., Meuschke, N., Aizawa, A., and Gipp, B. Identification and Analysis of Media Bias in News Articles. *Proceedings of the 15th International Symposium of*

Information Science, (2017).

[20] Hussain, J., Khan, W.A., Afzal, M., Hussain, M., Kang, B.H., and Lee, S. Adaptive User Interface and User Experience Based Authoring Tool for Recommendation Systems. In *Ubiquitous Computing and Ambient Intelligence. Personalisation and User Adapted Services*. Springer, 2014, 136–142.

[21] Leutenegger, S., Chli, M., and Siegwart, R.Y. BRISK: Binary robust invariant scalable keypoints. *Computer Vision (ICCV), 2011 IEEE International Conference on*, (2011), 2548–2555.

[22] Lowe, D.G. Object recognition from local scale-invariant features. *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, (1999), 1150–1157.

[23] McNee, S.M., Albert, I., Cosley, D., et al. On the Recommending of Citations for Research Papers. *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, (2002).

[24] Merton, R.K. *The sociology of science: Theoretical and empirical investigations*. University of Chicago press, 1973.

[25] Meuschke, N. and Gipp, B. Reducing Computational Effort for Plagiarism Detection by using Citation Characteristics to Limit Retrieval Space. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries (DL14)*, (2014), 197–200.

[26] Meuschke, N., Gipp, B., and Breitinger, C. CitePlag: A Citation-based Plagiarism Detection System Prototype. *Proceedings of the 5th International Plagiarism Conference*, (2012).

[27] Meuschke, N., Gondeck, C., Seebacher, D., Breitinger, C., Keim, D., and Gipp, B. An Adaptive Image-based Plagiarism Detection Approach. *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, (2018).

[28] Meuschke, N., Schubotz, M., Hamborg, F., Skopal, T., and Gipp, B. Analyzing Mathematical Content to Detect Academic Plagiarism. *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, (2017).

[29] Meuschke, N., Siebeck, N., Schubotz, M., and Gipp, B. Analyzing Semantic Concept Patterns to Detect Academic Plagiarism. *Proceedings International Workshop on Mining Scientific Publications (WOSP) held in conjunction with the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, (2017).

[30] Mihalcea, R., Corley, C., Strapparava, C., and others. Corpus-based and knowledge-based measures of text semantic similarity. *AAAI*, (2006), 775–780.

[31] Mutlu, B., Veas, E.E., Trattner, C., and Sabol, V. VizRec: A Two-Stage Recommender System for Personalized Visualizations. *IUI*, (2015).

[32] Pagael, R. and Schubotz, M. Mathematical Language Processing Project. *arXiv preprint arXiv:1407.0167*, (2014).

[33] Reiterer, H., Tullius, G., and Mann, T.M. INSYDER: a content-based visual-information-seeking system for the web. *International Journal on Digital Libraries 5*, 1 (2005), 25–41.

[34] Schubotz, M. *Augmenting mathematical formulae for more effective querying & efficient presentation*. 2017.

[35] Schubotz, M., Grigorev, A., Leich, M., et al. Semantification of Identifiers in Mathematics for Better Math Information Retrieval. *energy 2*, 5 (2016), 10.

[36] Sinha, R.R. and Swearingen, K. The role of transparency in recommender systems. *CHI*, (2002).

[37] Tang, T. and McCalla, G. Utilizing artificial learners to help overcome the cold-start problem in a pedagogically-oriented paper recommendation system. *International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, (2004), 245–254.

[38] Veltkamp, R.C., Burkhardt, H., and Kriegel, H.-P. *State-of-the-art in content-based image and video retrieval*. Springer Science & Business Media, 2013.

[39] Voorhees, E.M. Query Expansion Using Lexical-Semantic Relations. *SIGIR*, (1994).

[40] Ware, M. and Mabe, M. The STM report: An overview of scientific and scholarly journal publishing. (2015). https://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf

[41] Woodruff, A., Gossweiler, R., Pitkow, J., Chi, E.H., and Card, S.K. Enhancing a Digital Book with a Reading Recommender. *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, ACM (2000).

[42] Zhang, J., Deng, B., and Li, X. Concept Based Query Expansion Using WordNet. *2009 International e-Conference on Advanced Science and Technology*, (2009), 52–55.