What Should I Cite? Cross-Collection Reference Recommendation of Patents and Papers

Julian Risch and Ralf Krestel

Hasso-Plattner-Institut, Prof.-Dr.-Helmert-Str. 2–3, 14482 Potsdam, Germany
firstname.lastname@hpi.de

Abstract. Research results manifest in large corpora of patents and scientific papers. However, both corpora lack a consistent taxonomy and references across different document types are sparse. Therefore, and because of contrastive, domain-specific language, recommending similar papers for a given patent (or vice versa) is challenging.
We propose a recommender system that leverages topic distributions and keywords to recommend related work despite these challenges. As a case study, we evaluate our approach on patents and papers of two fields: medical and computer science. We find that topic-based recommenders complement word-based recommenders for documents with collection-specific language and increase mean average precision by up to 27%. As a result of our work, publications from both corpora form a joint digital library, which connects academia and industry.

Keywords: Recommender systems, Text mining, Topic modeling

1 Searching for Related Work across Patents and Papers

More than 1.2 million patents will be granted[1] and more than 1.5 million scientific papers will be published in 2017 according to bibliometric growth models [1]. These large collections form an extensive library of latest research results in an almost unstructured form, thus challenging to mine automatically. Searching for related work in papers is an important task for academic researchers. Similarly, patent applicants search for prior art to prove novelty and to define scope. Prior art denotes publicly available, state-of-the-art information in any form. Therefore, it is not limited to patents but includes also papers. Content-based recommender systems for text documents typically rely on tf-idf-based measures to identify representative keywords of a document and to recommend similar documents. However, linguistic differences of patents and papers are challenging for word-based recommender systems: Although a patent and a paper deal with the same topic, they might use different words to describe their work.

Because patents claim the scope of an invention, they cover as much variation of the invention as possible. As a consequence, patent descriptions use vague language, such as "electronic imaging apparatus", whereas a paper might

[1] http://www.wipo.int/edocs/pubdocs/en/wipo_pub_941_2016.pdf

call the same invention “digital camera”. Moreover, patents have specific linguistic characteristics, such as a higher frequency of words with indefinite, general meaning. Approximately 1% of scientific papers cite at least one patent [3]. Because existing references across patents and papers are sparse, we assume that these references are not suited to train graph-based recommender systems.

With this work, we propose cross-collection topic modeling to bridge the linguistic gap between patents and scientific papers. Based on topic distributions, we identify and recommend topically similar patents and papers even if they do not share keywords. In contrast to manual classification with inconsistent taxonomies, topic modeling is an unsupervised machine learning technique. As a consequence, our approach allocates topics to millions of documents automatically. We present two case studies on datasets consisting of U.S. patents, computer science papers, and medical articles. For an evaluation on these datasets, we use existing references as a gold standard for recommendations and compare the mean average precision (MAP) of topic-based, word-based, and combined recommender systems.

2 Related Work

Mining Patents and Papers. More than 200 research articles address recommender systems for scientific papers. For example, Liu et al. mine citation graphs of computer science literature to predict further citations [5]. Most recently, Momeni et al. evaluate how co-authorship networks support author name disambiguation for common names [8]. Wang et al. identify topics in patents based on latent Dirichlet allocation (LDA) and noun phrase extraction [10]. They compare different institutions with regards to their patents’ topic distributions. Krestel et al. propose a recommender system for patents based on topic modeling and document ranking techniques [4]. Although patent mining and paper mining face similar challenges, such as keyword extraction and topic modeling, they form two separate research fields. Especially different document style and the variation of wording limits the capabilities of holistic approaches. For example, Google Scholar[2] provides a search interface for patents and papers, but its word-based approach neglects linguistic contrasts.

Topic Modeling. Wang et al. combine collaborative filtering and topic modeling to recommend scientific papers in a user’s field of interest [11]. Given a citation graph, Mei et al. propose a concept of “topical inheritance” and enforce similar topic distributions in cited and citing documents [7]. However, all previous approaches consider only single collections and neglect linguistic contrasts of patents and papers. Extending LDA, Paul et al. model topics across multiple corpora with cross-collection LDA (ccLDA) [9]. Their approach considers collection-specific and collection-independent word distributions per topic but not in the domain of recommender systems or patents and scientific papers.

[2] `https://scholar.google.com`

Cross-Domain Recommendation. To match query terms and document terms in heterogeneous digital libraries, Mayr et al. propose to manually map terminology from one controlled vocabulary to another [6]. However, this approach requires an enormous manual effort. With our cross-collection topic model, we automate the matching of collection-specific and collection-independent terms. Recently, recommender systems have been proposed to transfer users' rating patterns from one domain to another, such as movies and books [2]. However, such cross-domain recommender systems rely on users' rating histories to transfer knowledge and our task lacks user ratings. To the best of our knowledge, so far no research addresses patents and scientific papers as a joint library of related work. Neither cross-collection topic models nor cross-domain recommender systems have been used to bridge the linguistic gap between both corpora.

3 Jointly Recommending Patents and Papers

To recommend similar patents or papers, we propose two complementing similarity measures based on (i) keywords and (ii) topic distributions. Whereas word-based similarity is fine-grained, topic-based similarity is coarser-grained.

Keyword Similarity. For each document, we extract 10 representative keywords[3] with highest tf-idf scores. The similarity of two documents is calculated based on this keyword vector representation. While keywords are an established relevance measure for document retrieval systems, such as Elasticsearch[4], they are constrained by the exact wording in a document. This limitation emerges as a problem on patents and papers, because they make intensive use of collection-specific language. Even closely related documents may not have any keywords in common.

Topic Distribution Similarity. To reveal documents with similar latent topics across both collections, we adapt the topic model ccLDA and distinguish patent-specific, paper-specific and collection-independent word distributions per topic. In contrast to Paul et al., we distinguish collection-specific and collection-independent word types instead of word tokens. Types with similar frequency in patents and papers are modeled with a single, collection-independent probability, whereas all other types are modeled with multiple, collection-specific probabilities. Because collection-specific and collection-independent word types together constitute a topic, even documents that have no words in common can share the same topic distribution. We train the adapted ccLDA model and estimate topic distribution, collection-specific word distributions, and collection-independent word distribution in 500 iterations[5]. To compare documents based on their topic distribution, we use cosine similarity.

[3] Larger keyword vectors increase runtime but do not improve result quality.

[4] `https://www.elastic.co/products/elasticsearch`

[5] parameters set as suggested in the original paper: $\beta = 0.01$, $\delta = 0.01$, $\gamma_1 = 1$, $\gamma_2 = 1$

Table 1. The number of documents and gold standard references per dataset

	#Patents	#Papers	#References
Computer Science Dataset	3,377	2,443	6,488
Medical Dataset	19,419	21,921	70,588

Word-Based and Topic-Based Recommender System. We propose a recommender system that leverages the best combination of these two similarity measures to rank and recommend related work across patents and papers. We transfer the concept of explicit relevance feedback in information retrieval systems, where users evaluate initial query results to control subsequent queries and improve relevance. In our scenario a patent applicant wants to retrieve relevant papers for his patent. Based on this patent, keyword-based recommendations are presented. If the user's information need is not fulfilled, the recommendation approach can be manually switched to topic-based recommendations. We do not rely on an automatically switching hybrid but on the explicit decision of the user.

4 Case Study

Our evaluation task is to recommend related papers for each patent in our datasets. Although this limited case study considers only inter-collection references from patents to papers, our approach works also for references from papers to patents as well as intra-collection references without any adjustments. We evaluate the mean average precision of the top 100 recommendations, MAP@100.

Datasets. The first dataset contains granted U.S. patents and referenced ACM papers. We extract patent abstracts from United States Patent and Trademark Office (USPTO) publications and ACM paper abstracts from a citation network dataset[6]. We assume that referenced documents are related work. Therefore, cross-collection references serve as the gold standard for recommendations in our evaluation. The second dataset is based on medical research projects funded by the National Institutes of Health (NIH). These projects are required to list their patent and paper publications in a public database[7]. We consider projects with at least one patent and one paper. Publications of the same project are assumed to be related work and therefore serve as the gold standard for reference recommendations. Table 1 lists the number of documents and cross-collection references for each dataset.

With a preliminary experiment for the topic-based approach, we determine the number of topics with the highest MAP@100 per dataset. To this end, we split the dataset by time: We determine the number of topics on the oldest 50% of the documents and use the most recent 50% for the final evaluation of

[6] https://bulkdata.uspto.gov/ and https://aminer.org/citation

[7] https://exporter.nih.gov/

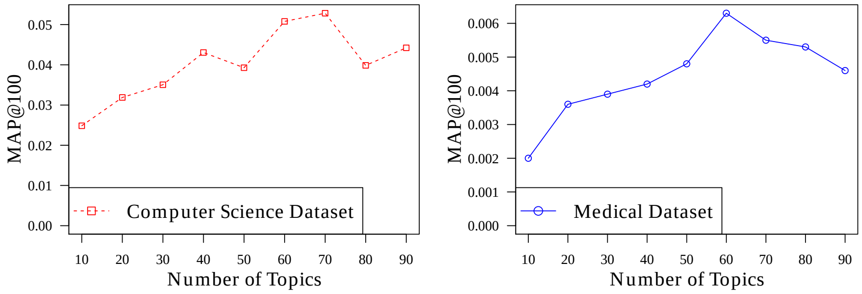

Fig. 1. MAP@100 of the topic-based approach for different numbers of topics

Table 2. MAP@100 comparison of topic-based, word-based, and combined approach

	Topic-Based	Word-Based	Best Comb.
Computer Science Dataset (70 Topics)	0.0528	0.1332	0.1696
Medical Dataset (60 Topics)	0.0068	0.0372	0.0414

MAP@100. According to the results visualized in Figure 1, we set the number of topics to 70 for the computer science dataset and to 60 for the medical dataset. We find that MAP@100 is consistently approximately one order of magnitude higher for the computer science dataset compared to the medical dataset. We assume, recommendation on the medical dataset is a more difficult task because of larger corpus size.

Recommendation Quality. Table 2 illustrates that the topic-based and the word-based approach are significantly outperformed by the best combination of both recommendation approaches on both evaluation datasets. Especially on the computer science dataset, the best combination achieves a 27% higher MAP@100 than the word-based approach. On the medical dataset, the MAP@100 is 11% higher. The experiment results demonstrate also that keywords are superior to topic distributions as a feature for recommending related work. However, their combination achieves the by far best results in our evaluation. Table 3 exemplifies a patent and its top three paper recommendations.

5 Conclusions and Future Work

In order to recommend patent and paper references despite their linguistic differences, we proposed a recommender system based on keywords and topic distributions of a cross-collection topic model. Experiment results demonstrate the effectiveness of this combination on two datasets of publications in the fields of medical and computer science. The combined approach outperforms word-based approaches by up to 27% MAP@100. A promising path for future work is to combine content-based and collaborative recommendation across patents and

Table 3. Top three recommendations for the patent "Method and apparatus for enhancing data storage efficiency". Relevant recommendations are in bold print.

Word-Based Paper Recommendations
1. Improving locality of reference in a garbage collecting memory management system
2. Garbage collection in a large LISP system
3. Page placement algorithms for large real-indexed caches
Topic-Based Paper Recommendations
1. A real-time garbage collector based on the lifetimes of objects
2. Garbage collection in a large LISP system
3. Design of the opportunistic garbage collector

papers. For example, authors could be compared based on their co-authorship relations or citation history. Furthermore, word-based and topic-based approaches could be combined for an automatic diversification of recommendations.

References

1. Bornmann, L., 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 (JASIST) 66(11), 2215–2222 (2015)
2. Gao, S., Luo, H., Chen, D., Li, S., Gallinari, P., Guo, J.: Cross-domain recommendation via cluster-level latent factor model. In: Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. pp. 161–176. Springer (2013)
3. Glänzel, W., Meyer, M.: Patents cited in the scientific literature: An exploratory study of 'reverse' citation relations. Scientometrics 58(2), 415–428 (2003)
4. Krestel, R., Smyth, P.: Recommending patents based on latent topics. In: Proc. of the Conf. on Recommender Systems (RecSys). pp. 395–398. ACM (2013)
5. Liu, Y., Niculescu-Mizil, A., Gryc, W.: Topic-link lda: Joint models of topic and author community. In: Proc. of the Int. Conf. on Machine Learning (ICML). pp. 665–672. ACM (2009)
6. Mayr, P., Mutschke, P., Petras, V.: Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. Library Review 57(3), 213–224 (2008)
7. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. In: Proc. of the Int. Conf. on World Wide Web (WWW). pp. 101–110. ACM (2008)
8. Momeni, F., Mayr, P.: Using co-authorship networks for author name disambiguation. In: Proc. of the Joint Conf. on Digital Libraries. pp. 261–262. ACM (2016)
9. Paul, M., Girju, R.: Cross-cultural analysis of blogs and forums with mixed-collection topic models. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP). pp. 1408–1417. ACL (2009)
10. Wang, B., Liu, S., Ding, K., Liu, Z., Xu, J.: Identifying technological topics and institution-topic distribution probability for patent competitive intelligence analysis: a case study in LTE technology. Scientometrics 101(1), 685–704 (2014)
11. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: Proc. of the Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD). pp. 448–456. ACM (2011)