Cross domain citation recommendation based on hybrid topic model and co-citation selection

Supaporn Tantanasiriwong*,
Sumanta Guha, and Paul Janecek,

Department of Computer Science and Information Management, Asian Institute of Technology,
P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand
Email: supaporn.ait@gmail.com
Email: guha@ait.asia
Email: paul.janecek@gmail.com

Choochart Haruechaiyasak,

National Science and Technology Development Agency,
111 Thailand Science Park (TSP), Phahonyothin Road,
Klong Nueng, Klong Luang, Pathumthani 12120, Thailand
Email: choochart.haruechaiyasak@nectec.or.th

Leif Azzopardi

University of Strathclyde
Department of Computer & Information Sciences
Livingstone Tower
26 Richmond Street
Glasgow, G11XH
Email: leif.azzopardi@strath.ac.uk

*Corresponding author

Abstract: Cross domain recommendations are of growing importance in the research community. An application of particular interest is to recommend a set of relevant research papers as citations for a given patent. This paper proposes an approach for cross-domain citation recommendation based on the Hybrid Topic Model and Co-Citation Selection. Using the topic model, relevant terms from documents could be clustered into the same topics. In addition, the Co-Citation Selection technique will help select citations based on a set of highly similar patents. To evaluate the performance, we compared our proposed approach with the traditional baseline approaches using a corpus of patents collected for different technological fields of biotechnology, environmental technology, medical technology and nanotechnology. Experimental results show our cross domain citation recommendation yields a higher performance in predicting relevant publication citations than all baseline approaches.

Keywords: cross domain recommender system; citation recommendation; cross domain citation recommendation; topic model; co-citation selection; information retrieval; keyphrase extraction tool; similarity measures; evaluation; ANOVA; analysis of variance.

1 Introduction

Nowadays, there is an overwhelming amount of information available online. Users find information they need using search engines. However, queries by keyword search tend to elicit large numbers of data items and most of the retrieved information is often not relevant to the user's interest. In addition, users having different vocabulary knowledge tend to have their own individual keyword usage patterns even when searching the same topic. As a result, conventional information retrieval techniques may fail to satisfy users with their immediate results. Moreover, it may take the user significant effort, subsequently, to scan the result set for useful items. Therefore, recommender systems have emerged to efficiently filter the data and suggest information which is closest to the user's requirement (Adomavicius and Tuzhilin, 2005)

In business, corporations seek competitive advantage over their rivals. Recommendation systems have become an indispensable tool for online businesses to satisfy their customers. For example, in the recommendation engine of Amazon.com[1], Linden et al., (2003) suggest new products to users in order to encourage their customers to buy more products. Netflix[2], another online business, provides the customer feedback from a movie

[1] http://www.amazon.com/.

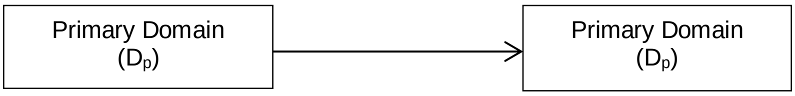
[2] http://www.netflix.com/.

recommendation system, which suggests video items that will likely interest the customer.

In general, there are two main approaches to information filtering in recommendation systems: collaborative filtering and content-based filtering (Adomavicius and Tuzhilin, 2005). First, the collaborative filtering approach allows matching an individual user with a group of users with similar preferences, and helps find items which the group, and hence the individual user, likely prefers. For example, in movie recommendation systems, the profiles of users who have similar preferences are collected and processed in the recommendation system to suggest movies based on prior ratings of users. Second, the content based filtering approach is to find items which match the user profile based on content characteristics. This technique is popularly applied in many areas such as online news (Claypool et al., 1999) , music (Liu and Tsai, 2001) , and web sites (Xu et al., 2005). The main problem of content based filtering is that we will only find items with a direct match to known characteristics, although users might actually be equally or more satisfied with items from other domains. The hybrid filtering approach, a combination of collaborative filtering and content-based filtering, aims to alleviate this problem by weighting the items and ranking the highest weights in order to suggest the appropriate items according to the user preference (Spiegel et al., 2009).

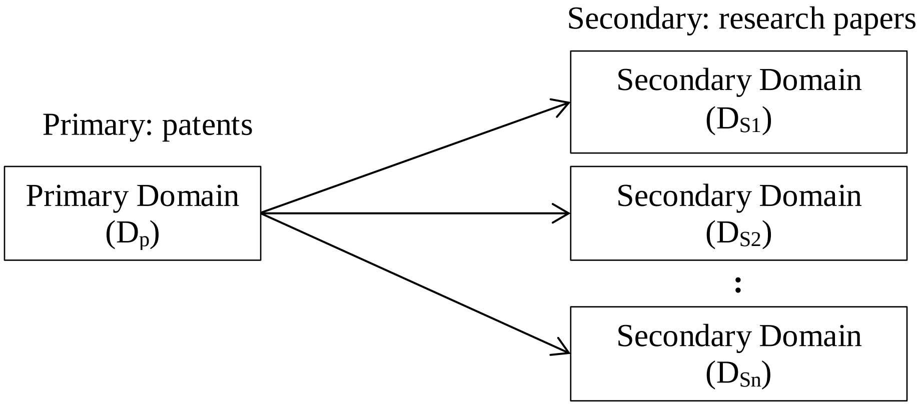
Recommender systems can be single or cross domain. Single domain recommender systems are those where the user's item ratings are processed within the latter's own domain. For instance, if the primary domain (D_p) is a set of books, then the books to be recommended are also derived from this domain. See Figure 1.

Figure 1 Single domain relationship



In a cross domain recommendation system, the recommended item is from a secondary domain. For example, a primary domain about patents (represented by D_p) can be used to suggest other research papers in a secondary domain (represented by D_{S1}, D_{S2},…, D_{Sn}). Figure 2 illustrate a

Figure 2 Cross domain relationship



cross domain relationship concept between a primary domain and a secondary domain for which patent documents and research papers are represented, respectively. These two collections may use different vocabularies, structures, and references reflecting the differences in the legal and academic research disciplines.

Currently, inventors looking for relevant existing patents will only find citations with in the domain collection, such as the academic literature domain (Strohman et al., 2007) and the patent retrieval domain (Fujii et al., 2007). Users rely on their own knowledge to search papers in each research paper database; for instance, the Institute of Electrical and Electronics Engineers[1] (IEEE), the Association of Computing Machinery[2] (ACM), and a service of the US National Library of Medicine[3] (PubMed). This problem inspired us to propose a Cross Domain Citation Recommender System (CDCRS) in order to help researchers gain useful recommendations of papers relevant to their research work across the patent and the research paper domains.

The remainder of this paper is organised as follows. Previous work related to cross domain and citation recommendation systems are described in Section 2. The relationship between a patent document and a research paper is illustrated in Section 3. Our proposed approaches to cross domain citation recommendation are described in Section 4 and 5 together with an implementation. Experiments and evaluation results are presented in Section 6. We conclude in Section 7.

[1] http://ieeexplore.ieee.org/.
[2] http://dl.acm.org/.
[3] http://www.ncbi.nlm.nih.gov/pubmed/.

2 Related works

In this section, existing works are first reviewed in terms of citation context where several methods are used to solve the citation recommendation problem. Next, in the second part, the initial works are described by technical usages in a variety of cross domain recommender systems.

Recommending citations for a manuscript usually relies on the information profile of the authors or the bibliography. For instance, McNee et al. (2002) conducted the collaborative filtering method for article recommendation and using citation network, paper citation, and co-citation information to perform a rating matrix based on the academic domain. The limitation of this paper is that they did not consider the content of the paper, which might help to select the appropriate papers for citation. Later, Hendrix (2005) solved the citation recommendation problem using a singular value decomposition (SVD) compared with collaborative filtering method. Strohman et al. (2007) proposed a combination of content features and citation graph features to measure the similarity between two documents for a citation recommendation system. They use Katz graph distance to rank a candidate set into the original set of documents. He et al. (2010) proposed a context-aware technique and probabilistic model to evaluate the relevance between documents and the citation contexts. He et al. (2011) proposed automatically recommending citation and identifying candidate citation contexts by examining the relevance segments between manuscript documents. Livne et al. (2014) focused on recommending citation to an academic paper using differential search. Lu et al. (2011) used a translation model for recommending citations by bridging languages from the citation contexts and the cited papers. They discovered that the context-aware relevance model was more effective than language modelling. But, the translation model outperformed both language model and context-aware model. Huang et al. (2012) considered a citation recommendation by adapting the translation model-based approach for mapping citation contexts with references. Tuarob et al. (2012) proposed a co-citation network algorithm, using the citeseer corpus, where graph based clustering is applied for linking documents and references. Liu et al. (2012) proposed the combination of PageRank and language model method in contrast to the baseline approaches TFIDF and BM25 for citation analysis based on Scientific Publication Collection. Su et al. (2009) focus on grouping reference papers from authors who publish more than one paper in order to find the authors who have multiple expertise based on co-citation analysis in the ACM journal domain. Therefore, the various techniques mentioned earlier are aimed at suggesting citations based on a discriminative, context-aware, translation model, and citation based graph network approaches predict the reference papers from their own manuscript.

In patent citation recommendations, Fujii et al. (2007) proposed citation analysis by combining text-based and citation-based scores to improve the invalidity search on patent retrieval. Rodriguez et al. (2015) proposed patent citation network analysis to identify the influence node of patents using a graph kernel measurement. Noh et al. (2015) focused on keyword selection and processing for patent analysis using factors of patent documents where the element and the number of selecting keyword, and transforming technique were considered to increase the reliability of this research. Generally, automatic keywords extraction from patent documents has been used in innovation management (Dou et al., 2005). Golestan Far et al. (2015) explores the term selection techniques of patent query in description section by integrating with BM25 and Language Model to upper bound state-of-the-art prior art search performance. Many researchers have focused on improving patent search retrieval by using various supervised and unsupervised learning techniques. For example, Verma and Varma (2011) compared supervised and unsupervised tools for invalidity search on patents and found that generating queries based on a Keyphrase Extraction Algorithm (KEA) as a supervised learning method performs better than the unsupervised approach.

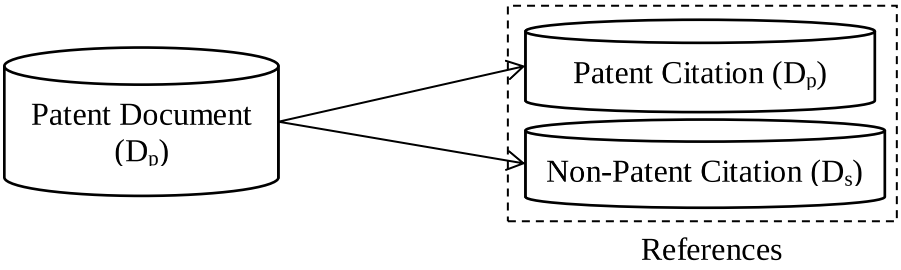
Some papers have attempted to integrate the WordNet (Varelas et al., 2005) lexical thesaurus to expand queries with synonymous terms (Zhang et al., 2009). Veeramachaneni (2010) focused on unsupervised learning for automatic re-ranking in patent retrieval. They used the WordNet vocabulary to enhance their thesaurus for the query expansion model. Tantanasiriwong and Haruechaiyasak (2013) used topic model expansion to adding relevant terms to reduce term mismatch between patents and citations. Our literature review on current research showed that most of the citation recommendation systems attempt to develop techniques based on a single source domain.

Knowledge is multidisciplinary, however our literature review found that most citation recommendation techniques are based on single source domain. Online bibliographic databases such as PubMed in the medical domain, IEEE in the engineering domain, ACM in the computing domain, and USPTO Patents in the innovation domain limit their search tools within their own collection. Finding relevant papers requires accessing each information source, as well as the appropriate domain specific vocabulary. Previous research has not addressed the particular needs of cross domain search. We also investigate the context of each domain and analyse co-citation relationship between two different domains, that is, patent and non-patent citation.

3 Patent document and research paper relationship

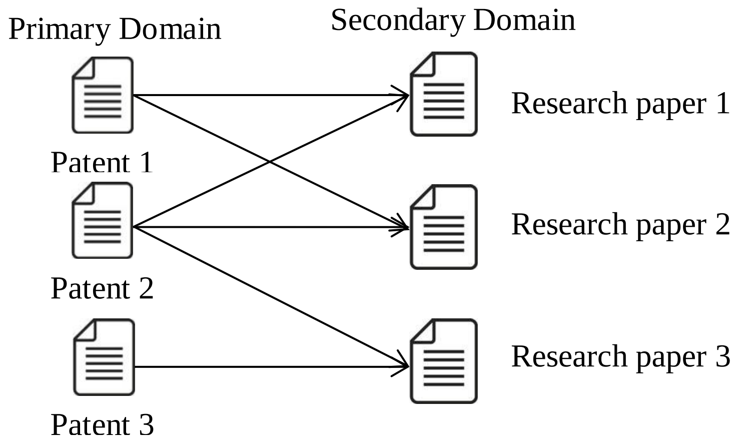
Patents play an important role in research and innovation. Commercially, they are legally protected by the laws of intellectual property. Most patent retrieval tools focus on using information retrieval to retrieve patent documents that satisfy the inventor's needs. However, to return the most relevant documents, it requires more sophisticated keyword inputs that overcome the limitation of knowledge in vocabulary of users. Therefore, we introduce a query by example algorithm using the patent document rather than standard keyword search methods. Patents contain several identifiable and independently important parts, including title, abstract, claim, description, summary, and references. The reference sections of a patent document consist of a patent citation section and a non-patent citation section. In this paper, the former section is called a primary domain (D_p), and the latter is called a secondary domain (D_s) as shown in Figure 3. Only patent citation papers can be found in the primary domain whereas the non-patent citation ones contain research papers from various domains such as IEEE or ACM, and so on. In our case, each domain is regarded as a separate information source.

Figure 3 A patent citation relationships



The diagram in Figure 4 illustrates a cross domain citation relationship between patents in primary domain and research papers in secondary domain, which are derived from their corresponding references of those patents in primary domain.

Figure 4 Example of cross domain citation relationship



The challenge here is the very different terminology used in patent and academic documents even though they are discussing the same topic. For example, an academic research paper may refer to a "router" where a patent document uses the term "gateway machine", as shown in Table 1.

Table 1 Words usage examples

Patent document	*Research paper*
Energy	Battery
Image Device	Camera
Memory	SRAM
Gateway Machine	Router

Notes: Each row shows examples of words usage in two document domains.

The purpose of this research is to develop an accurate and effective Cross Domain Citation Recommender System (CDCRS) to solve the problem of the cross domain citation for patent recommendation. Our contributions include a Hybrid Topic Model and Co-Citation Selection to resolve the cross domain citation recommendations. In cross domain recommendation, both patents and research papers are linked on the basis of two concepts. Firstly, a topic model concept, Topic Model-Based Reduction (TM-BR), is applied to reduce the dimensionality of patent documents. Secondly, a linkage concept is implemented such that there is a co-linking relation between a patent and its citations, called Co-Citation Selection (CCS). The CCS is implemented under the concept that patents with similar co-citations can be represented as a similar pair of patents. In Figure 4, for instance, Patent1 has more similarity to Patent 2 than to Patent3, because both Patent 1 and Patent 2 have a similar set of co-citations. In addition, our experiment was conducted to verify that a Hybrid Topic Model and Co-Citation Selection (HTC) framework work more effectively than any baseline method.

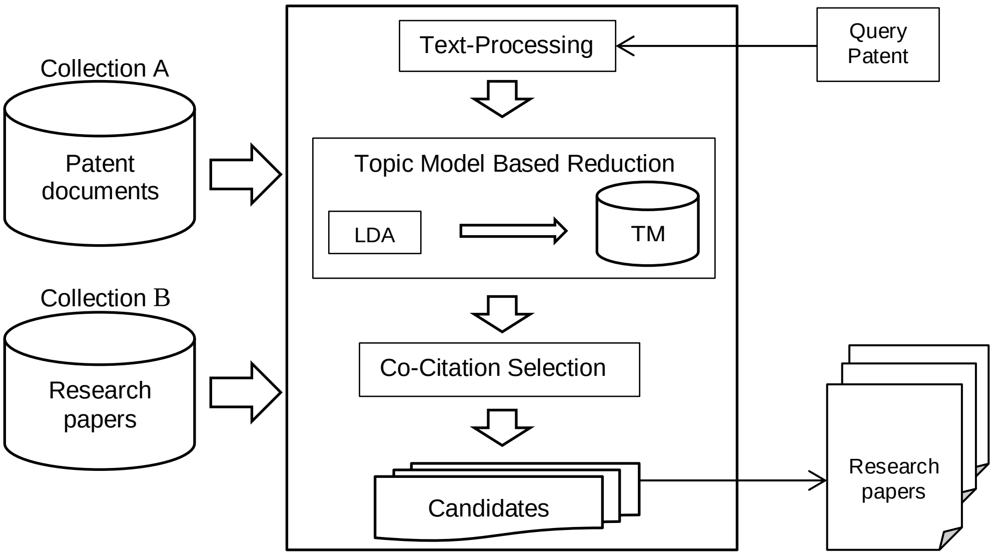
4 Cross Domain Citation Recommender System

In this section, we present the Cross Domain Citation Recommender System (CDCRS) by recommending research papers for a given patent document. The Hybrid of Topic Model combined with Co-Citation Selection (HTC) approach is proposed to improve the performance of the cross domain citation recommender system.

4.1 Hybrid of Topic Model and Co-Citation Selection (HTC)

The section describes a new framework for Cross Domain Citation Recommendation (CDCR) using a Hybrid of Topic Model-Based Reduction and Co-Citation Selection (HTC) as shown in Figure 5.

Figure 5 Cross Domain Citation recommendation framework using HTC



The proposed framework has three main steps; text-processing, topic model reduction, and finally co-citation selection. The first step extracts keywords and key phrases from the patent collection and research paper citation collection using the Maui-indexer (Medelyan et al., 2009), an extension of the standard key phrase extraction (KEA) algorithm. The second step generates the Topic Models for a patent query document using the Latent Dirichlet Allocation (LDA) algorithm, where a list of words in the patent query is represented by a list of topics. The third step generates the research paper citation for a particular patent. The following sections describe each step in greater detail.

- Topic Model

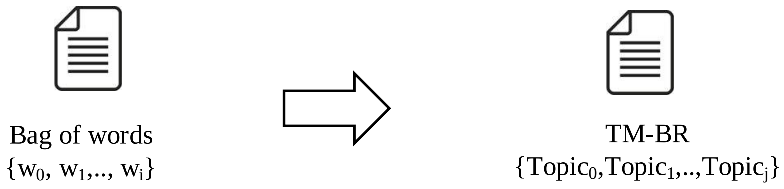
The Topic Model represents topics as a probability distribution over words (Steyvers and Griffiths, 2007) based on the LDA algorithm and the Gibbs sampling methods proposed by Blei et al. (2003). Equation 4 describes the parameters of a Topic Model:

$$P(w_i \mid d) = \sum_{j=1}^{z} P(w_i \mid z_i = j) P(z_i = j \mid d) \qquad \textbf{(1)}$$

where $P(w_i \mid d)$ is the probability of an arbitrary word w_i given by a document d, and z_i represents a latent topic over word distribution in a given document.

To recommend a collection of research paper citations for a given patent, a set of patent documents in a collection is given and denoted by $C = \{d_0, d_1, \ldots d_{i-1}\}$, where each patent document, d_i consists of a list of words denoted by $d_i = \{w_0^i, w_1^i, \ldots, w_{|d_i|-1}^i\}$ and $|d_i|$ is the total number of words in d_i.

Figure 6 The example of Topic Model-Based Reduction

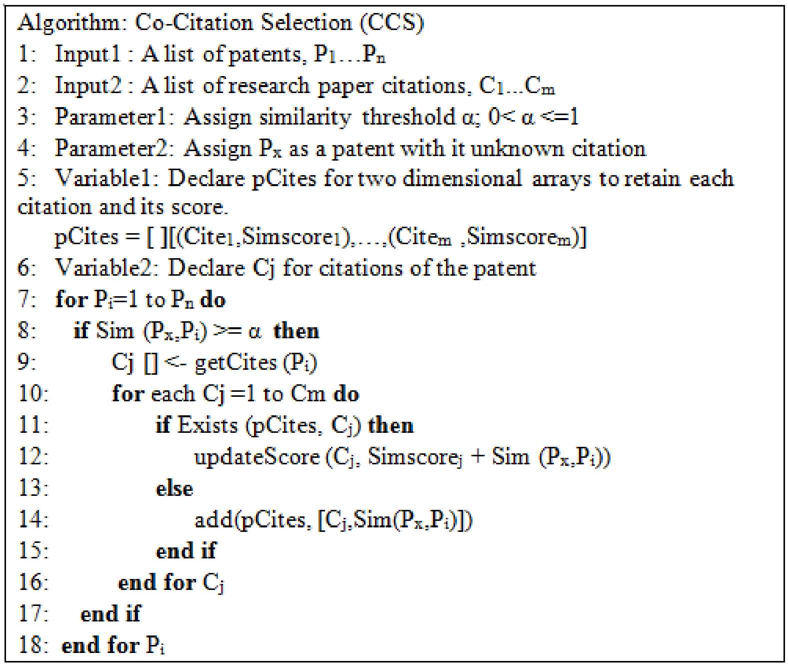


Topic Model-Based Reduction (TM-BR) represents documents as a probability distribution over a set of topics. As shown in Figure 6, a list of patent documents denoted by $d_i = \{w_0, w_1, ..., w_i\}$are transformed into new Topic Model representations of patent documents denoted by $d_j^{\prime} = \{T_0^j, T_1^j, ..., T_{m-1}^j\}$ as a set of topics where $T_k^j (k = 0, \ldots, m-1)$ is a probability of topic distribution in each patent document.

- Co-Citation Selection (CCS)

In this section, we give the details of the CCS algorithm as shown in Figure 7. The goal of this algorithm is to find research paper citations for a patent. To accomplish this task, we compare the new patents with the existing patent documents whose citations are already known. The candidates' research papers citations to a new patent can be generated by the CCS algorithm (Tantanasiriwong and Haruechaiyasak, 2014). This approach operates under the assumption that patents with similar context tend to have a similar set of citations. To start the algorithm, we first assign the similarity threshold (alpha) as a criterion to filter out a patent and its related citation whose similarity score is below the defined threshold. Then, we assign P_x parameter to represent a query patent with its unknown citation. Each citation of P_x will be reserved as the answer for subsequent evaluation. After that, the similarity between P_x and neighbouring patents is computed to find the score of relevant patents for P_x using the cosine similarity metric (Huang, 2008). At the end, ranking is carried out among those citations for patent-citations prediction.

Figure 7 The Co-Citation Selection (CCS) algorithm



```
Algorithm: Co-Citation Selection (CCS)
Input1 : A list of patents, $P_1 \ldots P_n$
Input2 : A list of research paper citations, $C_1 \ldots C_m$
Parameter1: Assign similarity threshold $\alpha$; $0 < \alpha <= 1$
Parameter2: Assign $P_x$ as a patent with it unknown citation
Variable1: Declare pCites for two dimensional arrays to retain each citation and its score.
    pCites = [ ][(Cite$_1$,Simscore$_1$),....,(Cite$_m$ ,Simscore$_m$)]
Variable2: Declare Cj for citations of the patent
for $P_i$=1 to $P_n$ do
    if Sim ($P_x$,$P_i$) >= $\alpha$ then
        Cj [] <- getCites ($P_i$)
        for each Cj =1 to Cm do
            if Exists (pCites, $C_j$) then
                updateScore ($C_j$, Simscore$_j$ + Sim ($P_x$,$P_i$))
            else
                add(pCites, [$C_j$,Sim($P_x$,$P_i$)])
            end if
        end for $C_j$
    end if
end for $P_i$
```

5 Experimental setup

5.1 *Data Collections and Pre-processing*

Two document sets were created for this evaluation: patent documents and research papers. The patent documents, as a primary document set, were collected from USPTO in four technology fields, in accordance with their International Patent Classification (IPC): Medical Technology, Biotechnology, Environment Technology, and Nanotechnology. The research papers, as a secondary document set, are typical of their patent documents' citation papers and were gathered and retrieved from IEEE publications.

Table 2 Summarisation of dataset collections

Category	*No. of primary domain (Patent)*	*No. of secondary domain (Publication)*	*No. of test documents*	*No. of unique keywords*
Medical	2,867	4,326	1,000	44,819
Biotechnology	2,229	3,697	1,000	44,063
Environmental	3,105	5,118	1,000	46,594
Nanotechnology	2,317	3,126	1,000	33,724

Notes: Each row shows a dataset collection and its corresponding sets of documents in different technology domains.

We prepared test documents by randomly sampling the documents that contains co-citations. Then, a new dataset collection was constructed that included 1,000 patents from 4 categories with their related research paper citations, as shown in Table 2. We extracted keywords and keyphrases for cross domain information using a tool called Maui, as recommended in (Tantanasiriwong et al., 2014). In addition, all documents were filtered by removing stop words[1]. The total number of unique words in patent-research papers obtained from keyphrase processing in each domain was as shown in Table 2: 44,819 medical; 44,063 biotechnologies; 46,594 environmental; and 33,724 nanotechnology.

5.2 Baseline approaches

To evaluate our proposed algorithms, we compare them with three baseline approaches. Two traditional models are presented: Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) based on the vector space model. And Best Match (BM25) model is introduced and applied as an average document length weight in each document.

The traditional information retrieval approach would be to discover the relevant documents based on keywords given in a user's queries (Blair, 1979). The vector space model (VSM) represents documents as a vector of the terms that occur in the document. In information retrieval, term

[1] http://www.lextek.com/manuals/onix/stopwords1.html/.

weighting within VSM is commonly represented as term frequency (TF) and term frequency-inverse document frequency (TF-IDF) (Salton and Buckley, 1988). Equations (1) and (2) define the general forms of TF and TF-IDF:

TF: $tf = 1 + \log(tf)$, where tf>0 **(2)**

TF-IDF: $tf * idf = 1 + \log(tf) * \log(\frac{N}{df})$ **(3)**

where N is the number of documents in the collection, and df is the number of documents where the term appears within the collection.

We processed both patent document and research paper and represented them as a term frequency vector prior to measuring the similarity of those two domains. However, these traditional approaches also can apply through the BM25 technique where the term weight is adjusted by BM25 score as follow.

BM25 or Best Match is a classic probabilistic model in information retrieval. The score of BM25 is computed using query keywords that appear in each document and the document length normalisation feature (Jones et al., 2000). Equation (3) defines the term weight in BM25:

$$Wi = \log\frac{N}{df(w_i)} * (\frac{f(w_i, D) * (k+1)}{f(w_i, D) + k(1 - b + b * \frac{|D|}{avgdl})}) \quad \textbf{(4)}$$

where $f(w_i, D)$ is term frequency of words in the document, $|D|$ is the length of each document, and *avgdl* is the average document length in the document collection. Here, N is the number of documents in the collection, and $df(w_i)$ is the number of documents in which w_i appears, and k = 0.5 and b = 0.75 are the constants assigned by user.

5.3 Evaluation Metrics

To evaluate the performance of each query in the testing set, we use Precision, Recall, F-Measure, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR) (Radev et al., 2002) as defined in equation (5),(6),(7),(8),(9) as follows.

- Precision is the fraction of retrieved citation documents that are relevant to the user query.

$$\text{Precision} = \frac{|Ra \cap Rr|}{|Rr|} \qquad \textbf{(5)}$$

where Ra is the set of relevant documents and Rr is the set of retrieved documents.

- Recall is the fraction of the relevant document and retrieved citation.

$$\text{Recall} = \frac{|Ra \cap Rr|}{|Ra|} \qquad \textbf{(6)}$$

where Ra is a relevant documents and Rr is a retrieved documents.

- F-Measure is an alternative solution for calculating the accuracy by considering both recall and precision.

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})} \qquad \textbf{(7)}$$

- Mean Average Precision (MAP) is the average precision across multiple queries. It considers only the rank position of each of the relevant documents and matches this to the query result item. The equation is as follows:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{q} \qquad \textbf{(8)}$$

where $AveP(q)$ is average precision in each query and q is the number of queries.

- Mean Reciprocal Rank (MRR) is a measure of the average of the reciprocal ranks of query results. It is derived from a list of results ordered by probability of correctness. The equation is as follows:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \qquad \textbf{(9)}$$

where, $|Q|$ is the number of testing queries.

6 Experiments and evaluation results

In our experiment, the cross domain citation matching technique is computed based on the standard cosine similarity. The four different approaches are appraised by performance metrics in Tables 3, 4 and

Figure 8. Each approach is evaluated, compared, and summarised as shown in Figure 9.

6.1 Baseline approaches

Three baseline methods are applied to this framework: TF, TF-IDF, and BM25. Subsequently, the calculation of the traditional similarity matching between the domain of patent documents and the domain of research papers is performed. Prior to such similarity calculation, those two domains were to be transformed into the same dimension. The baseline result shows that the BM25 weighting method outperforms any other simple approach including the TF and TF-IDF techniques in all four categories based on the indexes of Mean Precision (MP), Mean Recall (MR), and Mean F-measure (MF) as shown in Table 3.

Table 3 The performance of baseline approaches

	Medical			*Biotechnology*			*Environment*			*Nanotechnology*		
	MP	*MR*	*MF*	*MP*	*MR*	*MF*	*MP*	*MR*	*MF*	*MP*	*MR*	*MF*
TF	0.01	0.42	**0.03**	0.02	0.33	**0.04**	0.02	0.39	**0.03**	0.02	0.42	**0.03**
TF-IDF	0.02	0.53	**0.03**	0.03	0.40	**0.04**	0.02	0.48	**0.04**	0.02	0.52	**0.04**
BM25	0.02	0.60	**0.04**	0.03	0.49	**0.05**	0.03	0.52	**0.05**	0.02	0.56	**0.04**

Notes: The best MF values for each method in each category are emphasized in bold.

6.2 Co-Citation Selection approach

In CCS, the effectiveness of CDCR is measured by Mean F-Measure in terms of neighbouring patent-selection and threshold adjustment. In table 4, the Mean F-Measure results indicate that the CCS approach achieved the highest accuracy with a Threshold (TH) Cutoff at 0.7. This phenomenon happens in all technology fields.

Table 4 The performance of CDCR based on CCS approach

TH	*Medical*			*Biotechnology*			*Environment*			*Nanotechnology*		
	MP	*MR*	*MF*	*MP*	*MR*	*MF*	*MP*	*MR*	*MF*	*MP*	*MR*	*MF*
0.3	0.03	0.63	0.05	0.03	0.46	0.06	0.03	0.55	0.06	0.03	0.45	0.05
0.5	0.05	0.57	0.08	0.04	0.44	0.06	0.05	0.50	0.08	0.03	0.42	0.05
0.7	0.08	0.46	**0.12**	0.05	0.40	**0.08**	0.07	0.40	**0.10**	0.04	0.38	**0.06**

Notes: The best MF values at particular thresholds are highlighted in bold its corresponding data category.

6.3 Hybrid Topic Model and Co-Citation Selection approach

In the HTC approach, we present the results by varying the number of topics from 100 (T100) to 600 (T600). The evaluation results show that HTC generated more effective results than previous approaches. The topic of T600 and threshold cut-off at 0.7 achieves the highest mean F-Measure score of 43% in Medical, 36% in Biotechnology, 38% in Environment, 37% in Nanotechnology as shown in Table 5, whereas the topic of T100 with threshold at 0.7 has the lowest scores of 1.7% in Medical, 1.4% in Biotechnology, 1.3% in Environment, and 1.4% in Nanotechnology.

Table 5 The performance of the HTC approach for CDCR in each category using Mean Precision (MP), Mean Recall (MR), and Mean F-Measure (MF).

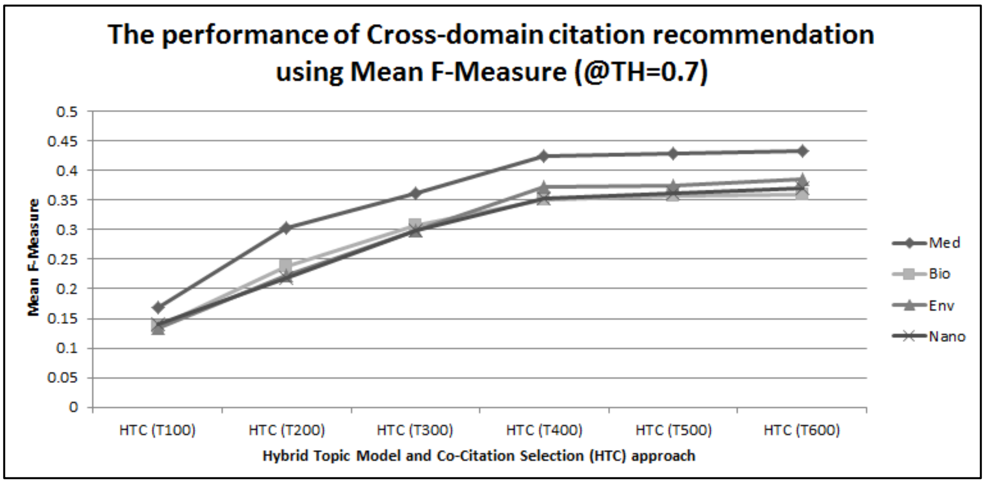
			Medical			*Biotechnology*			*Environment*			*Nanotechnology*	
Topic	*CCS*	*MP*	*MR*	*MF*	*MP*	*MR*	*MF*	*MP*	*MR*	*MF*	*MP*	*MR*	*MF*
T100	0.3	0.09	0.66	0.14	0.08	0.60	0.12	0.07	0.58	0.11	0.07	0.60	0.11
	0.5	0.10	0.64	0.16	0.09	0.58	0.13	0.08	0.55	0.13	0.08	0.59	0.13
	0.7	0.11	0.61	**0.17**	0.09	0.57	**0.14**	0.09	0.54	**0.13**	0.09	0.57	**0.14**
T200	0.3	0.21	0.65	0.29	0.16	0.57	0.22	0.15	0.57	0.20	0.14	0.58	0.19
	0.5	0.22	0.62	0.30	0.18	0.55	0.23	0.16	0.54	0.21	0.15	0.55	0.21
	0.7	0.23	0.61	**0.30**	0.18	0.54	**0.24**	0.17	0.53	**0.22**	0.16	0.53	**0.22**
T300	0.3	0.29	0.60	0.35	0.23	0.55	0.28	0.23	0.53	0.28	0.22	0.56	0.27
	0.5	0.30	0.57	0.36	0.25	0.53	0.30	0.24	0.51	0.29	0.24	0.53	0.29
	0.7	0.30	0.55	**0.36**	0.26	0.52	**0.31**	0.25	0.50	**0.30**	0.25	0.52	**0.30**
T400	0.3	0.37	0.59	0.42	0.30	0.53	0.35	0.32	0.53	0.36	0.28	0.55	0.33
	0.5	0.38	0.56	0.42	0.31	0.51	0.35	0.33	0.51	0.37	0.30	0.52	0.35
	0.7	0.38	0.55	**0.42**	0.31	0.50	**0.35**	0.33	0.50	**0.37**	0.31	0.50	**0.35**
T500	0.3	0.40	0.57	**0.44**	0.31	0.51	0.34	0.34	0.51	0.37	0.31	0.50	0.34
	0.5	0.40	0.55	0.43	0.32	0.49	0.35	0.34	0.49	0.37	0.32	0.48	0.36
	0.7	0.39	0.53	0.43	0.33	0.48	**0.36**	0.35	0.48	**0.37**	0.33	0.47	**0.36**
T600	0.3	0.42	0.55	**0.45**	0.33	0.48	0.36	0.36	0.49	**0.39**	0.33	0.49	0.36
	0.5	0.42	0.52	0.44	0.33	0.45	0.36	0.37	0.47	0.39	0.34	0.46	0.37
	0.7	0.41	0.50	0.43	0.34	0.44	**0.36**	0.36	0.46	0.38	0.35	0.45	**0.37**

Notes: The best MF values at particular CCS thresholds are highlighted in bold for each Topic Model in differrent categories.

In Figure 8, the line graph shows the performance of the HTC approach in CDCR over topics by ranking the number of topics. Increasing the number of topics in the experiment increases the Mean F-Measure value.

We also found that TH=0.7 performs better than any other TH for all field categories.

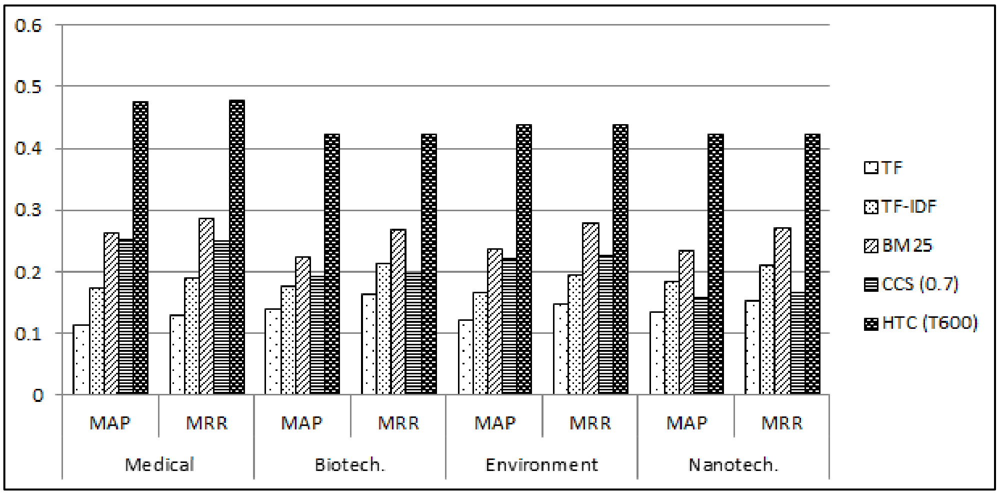
Figure 8 Performance evaluation of HTC approach in CDCR over topics based on F-Measure in difference fields of innovation



7 Comparative analysis of algorithms

In Figure 9, the bar chart shows performance comparisons of the cross domain citation recommendation using MAP and MRR based on the following approaches: TF, TF-IDF, BM25, CCS, and HTC.

Figure 9 Performance Comparison of five approaches using MAP and MRR



Statistical significance tests based on Analysis of Variance (ANOVA) were performed by using Least Significant Difference (LSD) to verify the effectiveness of each approach as shown in Table 6. The research hypothesis of these algorithms is that there are differences between means of average precision. The one-way ANOVA result shows that the means of BM25, CCS and HTC are 0.35, 0.49 and 0.90, respectively. In which, Mean of Average Precision (MAP) has statistically difference at 95% level of confidence among different algorithms.

Table 6 Comparison of Algorithms using ANOVA

Algorithm	N	Mean	Std. Deviation
BM25	683	0.35	0.330
CCS	449	0.49	0.400
HTC	490	0.90	0.218
Total	1622	0.55	0.398

Notes: Each row shows the ANOVA description in different algorithms.

Therefore, we accept the research hypothesis with the differences among these three algorithms. The significance test result demonstrates that all pairs of three algorithms are statistically different in term of significant for (p-value < 0.0005). This indicates that HTC achieves a significantly higher mean value than CCS and BM25. Moreover, the HTC approach achieves the highest mean of average precision in cross domain citation recommendation.

8 Conclusion

In this paper, we proposed a novel cross-domain citation recommendation framework for identifying relevant documents in a target domain given an example document in a source domain. The framework relies on a Hybrid Topic Model and Co-Citation Selection (HTC) algorithm. We evaluated this framework with a case study of patents and research articles. Our study showed that patents transformed using a topic model-based reduction and then integrated into CCS supports finding bibliographic information across domains. We compared the HTC approach with four baseline approaches (TF, TF-IDF, BM25, and CCS), and found that the HTC performs significantly better than the baseline approaches for cross domain citation recommendation.

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', *Knowledge and Data Engineering, IEEE Transactions on*, Vol. 17, No. 6, pp. 734-749.

Blair, D.C. 1979, *Information Retrieval, CJ Van Rijsbergen. London: Butterworths*, Wiley Online Library.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) 'Latent dirichlet allocation', *the Journal of machine Learning research*, Vol. 3, pp. 993-1022.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D. and Sartin, M. (1999), 'Combining content-based and collaborative filters in an online newspaper' in *Proceedings of ACM SIGIR workshop on recommender systems*, Citeseer.

Dou, H., Leveillé, V., Manullang, S. and Dou Jr, J.M. (2005) 'Patent analysis for competitive technical intelligence and innovative thinking', *Data science journal*, Vol. 4, pp. 209-236.

Fujii, A., Iwayama, M. and Kando, N. (2007), 'Overview of the patent retrieval task at the NTCIR-6 workshop' in *Proceedings of the Sixth NTCIR Workshop Meeting*, pp. 359-365.

Golestan Far, M., Sanne, S., Bouadjenek, M.R., Ferraro, G. and Hawking, D. (2015), 'On Term Selection Techniques for Patent Prior Art Search' in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 803-806.

He, Q., Kifer, D., Pei, J., Mitra, P. and Giles, C.L. (2011), 'Citation recommendation without author supervision' in *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, pp. 755-764.

He, Q., Pei, J., Kifer, D., Mitra, P. and Giles, L. (2010), 'Context-aware citation recommendation' in *Proceedings of the 19th international conference on World wide web*, ACM, pp. 421-430.

Hendrix, T. (2005) 'Can't See the Forest for the Trees?', *Cal Poly Magazine*, Vol. 9, No. 3, p. 7.

Huang, A. (2008), 'Similarity measures for text document clustering' in *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, pp. 49-56.

Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C.L. and Rokach, L. (2012), 'Recommending citations: translating papers into references' in *Proceedings of the 21st ACM international* *conference on Information and knowledge management*, ACM, pp. 1910-1914.

Jones, K.S., Walker, S. and Robertson, S.E. (2000) 'A probabilistic model of information retrieval: development and comparative experiments: Part 1', *Information Processing & Management*, Vol. 36, No. 6, pp. 779-808.

Linden, G., Smith, B. and York, J. (2003) 'Amazon. com recommendations: Item-to-item collaborative filtering', *Internet Computing, IEEE*, Vol. 7, No. 1, pp. 76-80.

Liu, C.-C. and Tsai, P.-J. (2001), 'Content-based retrieval of mp3 music objects' in *Proceedings of the tenth international conference on Information and knowledge management*, ACM, pp. 506-511.

Liu, X., Zhang, J. and Guo, C. (2012), 'Full-text citation analysis: enhancing bibliometric and scientific publication ranking' in *Proceedings of the 21st ACM international conference on Information and knowledge management*, ACM, pp. 1975-1979.

Livne, A., Gokuladas, V., Teevan, J., Dumais, S.T. and Adar, E. (2014), 'CiteSight: supporting contextual citation recommendation using differential search' in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, pp. 807-816.

Lu, Y., He, J., Shan, D. and Yan, H. (2011), 'Recommending citations with translation model' in *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM, pp. 2017-2020.

McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A. and Riedl, J. (2002), 'On the recommending of citations for research papers' in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, ACM, pp. 116-125.

Medelyan, O., Frank, E. and Witten, I.H. (2009), 'Human-competitive tagging using automatic keyphrase extraction' in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, Association for Computational Linguistics, pp. 1318-1327.

Noh, H., Jo, Y. and Lee, S. (2015) 'Keyword selection and processing strategy for applying text mining to patent analysis', *Expert Systems with Applications*, Vol. 42, No. 9, pp. 4348-4360.

Radev, D.R., Qi, H., Wu, H. and Fan, W. (2002) 'Evaluating web-based question answering systems', *Ann Arbor*, Vol. 1001, p. 48109.

Rodriguez, A., Kim, B., Lee, J.-M., Coh, B.-Y. and Jeong, M.K. (2015) 'Graph kernel based measure for evaluating the influence of patents in a patent citation network', *Expert Systems with Applications*, Vol. 42, No. 3, pp. 1479-1486.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information processing & management*, Vol. 24, No. 5, pp. 513-523.

Spiegel, S., Kunegis, J. and Li, F. (2009), 'Hydra: a hybrid recommender system [cross-linked rating and content information]' in *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, ACM, pp. 75-80.

Steyvers, M. and Griffiths, T. (2007) 'Probabilistic topic models', *Handbook of latent semantic analysis*, Vol. 427, No. 7, pp. 424-440.

Strohman, T., Croft, W.B. and Jensen, D. (2007), 'Recommending citations for academic papers' in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 705-706.

Su, Y.-M., Yang, S.-C., Hsu, P.-Y. and Shiau, W.-L. (2009) 'Extending co-citation analysis to discover authors with multiple expertise', *Expert Systems with Applications*, Vol. 36, No. 3, pp. 4287-4295.

Tantanasiriwong, S. and Haruechaiyasak, C. (2013), 'Patent Citation Recommendation Based on Topic Model Expansion' in *The Second Asian Conference on Information Systems, ACIS 2013*, pp. 316-319.

Tantanasiriwong, S. and Haruechaiyasak, C. (2014), 'Cross-domain citation recommendation based on Co-Citation Selection' in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2014 11th International Conference on*, IEEE, pp. 1-4.

Tantanasiriwong, S., Haruechaiyasak, C. and Guha, S. (2014) 'A Comparative Study of Key Phrase Extraction for Cross-Domain Document Collections', *The Emergence of Digital Libraries–Research and Practices*, Springer, pp. 393-398.

Tuarob, S., Mitra, P. and Giles, C.L. (2012), 'Improving algorithm search using the algorithm co-citation network' in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, ACM, pp. 277-280.

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E.G.M. and Milios, E.E. (2005), 'Semantic similarity methods in wordNet and their application to information retrieval on the web' in *Proceedings of the 7th annual ACM international workshop on Web information and data management*, ACM, Bremen, Germany, pp. 10-16.

Veeramachaneni, S. (2010), 'Unsupervised learning for reranking-based patent retrieval' in *Proceedings of the 3rd international workshop on Patent information retrieval*, ACM, pp. 23-26.

Verma, M. and Varma, V. (2011), 'Applying key phrase extraction to aid invalidity search' in *Proceedings of the 13th International Conference on Artificial Intelligence and Law,* ACM, pp. 249-255.

Voorhees, E.M. (1999), 'The TREC-8 Question Answering Track Report' in *TREC,* pp. 77-82.

Xu, B., Zhang, M., Pan, Z. and Yang, H. (2005) 'Content-based recommendation in e-commerce', *Computational Science and Its Applications–ICCSA 2005,* Springer, pp. 946-955.

Zhang, J., Deng, B. and Li, X. (2009) 'Concept Based Query Expansion Using WordNet', pp. 52-55.