Query-Driven Mining of Citation Networks for Patent Citation Retrieval and Recommendation

Parvaz Mahdabi, Fabio Crestani
University of Lugano, Faculty of Informatics,
Lugano, Switzerland
{parvaz.mahdabi, fabio.crestani}@usi.ch

ABSTRACT

Prior art search or recommending citations for a patent application is a challenging task. Many approaches have been proposed and shown to be useful for prior art search. However, most of these methods do not consider the network structure for integrating and diffusion of different kinds of information present among tied patents in the citation network. In this paper, we propose a method based on a time-aware random walk on a weighted network of patent citations, the weights of which are characterized by contextual similarity relations between two nodes on the network. The goal of the random walker is to find influential documents in the citation network of a query patent, which can serve as candidates for drawing query terms and bigrams for query refinement. The experimental results on CLEF-IP datasets (CLEF-IP 2010 and CLEF-IP 2011) show the effectiveness of encoding contextual similarities (common classification codes, common inventor, and common applicant) between nodes in the citation network. Our proposed approach can achieve significantly better results in terms of recall and Mean Average Precision rates compared to strong baselines of prior art search.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Retrieval Models, Query Formulation

General Terms

Experimentation, Performance, Measurement

Keywords

Prior art Search; Citation Graph; Bibliographic Network

1. INTRODUCTION

Patent prior art search is the task of recommending patent and non-patent documents which describe prior art work related to a patent application (referred to as query patent in this paper) and to provide a ranking of such relevant documents for the user. One of the main challenges in prior art search is related to the overwhelming term-mismatch. This problem originates from the frequent intentional obfuscation of content by patent writers and it leads to low retrieval effectiveness of patent search systems [9]. For example one patent document may contain few or no keywords in common with the query patent, while the idea conveyed in the two patent documents is quite similar. The question we try to answer in this paper is the following: given a time-evolving patent citation network, can we use the context of the query to overcome the term-mismatch problem and provide a ranking of relevant documents for a given patent application? We define the contextual information according to the type of information available between two linked patent documents on the citation network, namely: number of common applicants, number of common inventors, number of common classification codes, lexical, and temporal similarity. We exploit such similarities among linked patents as complementary information to the initial query in order to overcome the term-mismatch problem.

Recently, a few researchers [18, 22, 20, 12, 13] have studied the problem of recommending patent citations for prior art search. The first group [18, 22, 20, 15] mines heterogeneous networks derived from interacting patent companies and inventors. This group relies mainly on the link analysis and prediction on the network structure and do not take into account the term distribution of the patent documents. The second group [12, 13] targets textual content analysis through modeling the language of the query patent and integrating the proximity information into the query model. However, none of the above works target simultaneously all sources of available information, namely: the textual content of a query patent and the contextual relations available among interacting patents on a time-evolving patent citation network. The focus of the current paper is to perform query refinement by taking advantage of the rich network structure (node and edge attributes) as well as the term distribution of the interacting patents.

We propose a graph-based representation for a query patent according to the top ranked documents obtained from an initial rank list and their citation links. This graph-based framework takes into account the context-based relations between patent documents. This framework models the dependency between node attributes through the network structure. Our proposed method selects the influential patent

CIKM'14, November 3–7, 2014, Shanghai, China.

ACM 978-1-4503-2598-1/14/11 ...$15.00.
http://dx.doi.org/10.1145/2661829.2661899

documents from the query-specific citation network by running a random walker on a graph structure. The term distribution of selected candidates is then used for drawing concepts (unigram and word level bigrams) to formulate an expansion query model. We then provide a ranking model to recommend patent citations according to a query patent (i.e., a patent application submitted to the system).

In order to consider the dynamic nature of the network and recognize the new influential nodes, which are added to the network but have not stayed long enough to accumulate sufficient links, we parameterize the random walk with a time factor. We do this by considering the temporal order of the nodes in the citation network. We discount the initial probability of selecting a node as the seed of the Page Rank algorithm according to some temporal decay factor.

We experimented with different types of edge strengths to find out the effect of each similarity metric on the accuracy of citation recommendation. We find that using classification codes (topical similarity) has the highest performance among the similarity functions when they are used separately. The performance increases as the combination of different node attributes is used together to denote the strength of an edge. This shows the importance of leveraging as many node attributes as possible.

At an abstract level, our proposed approach is composed of three steps.

- **Generation of the citation network:** we extract the potential set of citation documents by performing similarity matching on lexical and topical levels for a query patent.
- **Network analysis:** where a time-aware random walk is used to identify the influential documents on the citation network and calculate a score for them.
- **Query refinement** the term distribution of candidate influential documents is used for drawing concepts (terms and word level bigrams). This query is used for performing a query expansion in order to re-rank the primary rank list.

Our contributions in this paper can be summarized as trying to answer the following questions:

- Are the citation network structure and diffusion of the contextual-based similarity between patent documents derived from the network structure useful for distinguishing between relevant and non-relevant patent documents?
- Can we use the network structure and the node attributes to better estimate a query model for performing patent retrieval?
- Does the combination of different node attributes for calculating edge weights improve over the result of using only a single node attribute?

We evaluate our framework on two large patent datasets, CLEF-IP 2010 and CLEF-IP 2011. We use the importance values associated to the nodes on the network to guide the query formulation process. This resembles pseudo relevance feedback and we perform such a task in a language modeling framework. We compare our proposed framework to the state of the art prior art search methods. We quantify the accuracy of our method in terms of the final ranking produced for recommending citations for a given patent application. Overall, our method achieves a significant improvement, $8-10\%$ in terms of recall@1000 and $8-20\%$ in terms of Mean Average Precision (MAP)@1000, over baseline methods on both datasets. We add different types of information to denote the strength of edges in the network such as similarity in terms of applicant, inventor, classification codes, lexical, and temporal information. We find that combining these weights increases the performance of the method. Our proposed framework is able to successfully use different types of nodes and edge attributes in order to recommend citations for a patent application.

Our paper is organized as follows: Section 2 reviews some related works. Section 3 describes the patent citation network and introduces one of the data sets and reports some observations we discovered while extracting different types of information from this dataset which motivates the usage of node information and edge attributes on the patent citation network. Section 4 introduces our initial query model estimated from the patent query. Section 5 explains our proposed model for time-aware network analysis and describes the algorithm for calculating node importance on the citation network. Section 6 explains the way we formulate a query by combining the information derived from the citation network with our initial query model. Section 7 reports our experiments that validate the effectiveness of our methodology, including its setup, baseline methods and evaluation results. Finally, Section 8 concludes this work.

2. RELATED WORK

We summarize the related work to this paper and arrange it along the two following dimensions. First, we focus on the proposed methods for prior art search. Second, we consider methods that take into account graph structure and model the dependency between nodes for patent citation recommendation.

Patent examiners apply term proximity heuristics in their searches in order to reward a document where matched query terms occur close to each other. This shows that proximity information plays an important role in real scenarios of patent searching. A few work [4, 12, 13, 2] studied the importance of proximity heuristics and phrases on patent search and classification. The work reported in [4] uses bigrams for patent classification. They obtained significant improvements by adding phrases. They concluded that bigrams are the most informative type of phrases for patent classification.

A recent study [13] builds a lexicon based on the patent classification code definitions[1]. This lexicon is used as a domain dependent resource for extracting specific terms to the topic

[1]See `http://web2.wipo.int/ipcpub/`

of the query. Proximity information is used to calculate reliable importance weights for the expansion concepts.

Another study on improving the retrievability of patent documents [2], combined term proximity heuristics with other features to select good terms for query expansion. In this work different distance functions were considered using proximity heuristics in comparison to standard query expansion.

A different approach is introduced in [20] which considers the co-invention relationships between inventors on the enterprise social network for recommending patent partners. They used a ranking factor graph model to predict future collaborators according to a user's profile and provide a recommendation list. They found that factors like, complementary research interests, and geographical proximity have positive effects on forming collaborations among inventors. The co-invention relationship is considered as an attribute in our patent citation network.

A recent work on patent citation recommendation [18], studies a heterogeneous network of patents including different type of objects such as companies, inventors and the technical content of patent documents. They identify the topical evolution of such objects on the patent network and provide a variety of micro level statistics to simplify the decision making of the user. For example, the system identifies active companies in the area and finds company's competitors according to their trend of technology development.

A different work [15] uses different type of features on a patent citation-bibliographic network and uses the RankSVM model to provide a rank list of citations for a given query patent. They achieve improvements at both precision and recall levels. In addition to the features used in this work, our current model also includes temporal information on the network.

Another related work [22] identifies competitive relationships among companies by learning across multiple heterogeneous networks. Authors study competitive relationship patterns on a company network, derived from a patent dataset and augmented by social networking information extracted from Twitter. They use topic modeling and build the topic model of each company, associating each company to a topic distribution. Their intuition is that entities with similar topic distributions are more likely to be competitors. They model the competitive relationship as a latent topic and use a factor graph model to infer the competitive label of each relationship among companies on the network. Experimental results show that their model is able to extract complementary competition patterns over these two sources, namely, the patent data set and the social network of Twitter.

It is worth mentioning that a number of researchers, [14] and [19], have looked into the problem of paper reviewers recommendation that is classified as expert finding. Patent citation finding and recommendation can also be seen as an instance of expert finding.

3. PATENT CITATION NETWORK

In this section we provide the problem definition and the motivation for our proposed model. We then discuss the motivation for representing the patent collection as a directed weighted graph and explain how to build it. We focus on the influence analysis aiming to find the important documents in the citation graph who could influence their domain terminology.

3.1 Problem Definition

The patent data set can be considered as a directed weighted graph $G = (V, E)$, where v is a set of $|V| = N$ patent documents and $E \in V \times V$ is a set of citation relationships between patent documents. Let x_i be a set of attributes associated with patent document v_i. An attribute can be the patent's inventors, patent's applicant, the classification codes (denoting the topic of the patent), and its publication date. We use $X = \{x_1, x_2, ..., x_N\}$ to denote the attributes of all patents. These attributes are later used in Section 5 to associate weights to edges. Our goal is to suggest a ranked list of citations for a specific patent application v_q, based on its attributes x_q, and the textual content of v_q.

3.2 Brief Introduction to the Data

CLEF-IP 2010 dataset is composed of over 2.6 million European patent documents corresponding to approximately 1.3 million individual patents published between 1985 and 2001. CLEF-IP patent documents are available[2] in XML format. We extracted the title, abstract, description, claims, patent publication date, and bibliographic data, such as classification codes, inventor names, and applicant names for all documents in the test set and the collection. We further extracted references (citations) for each patent in the data set excluding queries (patent applications in the test set).

It is worth mentioning that the relevance judgments for the CLEF-IP challenge are built using the documents listed in a search report for a patent application. The search report is written by the patent examiner. This report might share references with the initial citation list provided by the patent applicant. To remove the bias that might be introduced by the applicant, we do not consider the initial citations of the query patent in our model.

3.3 Motivation

A major motivation for our work comes from an inspection performed on the data set to test whether the contextual similarity between patent documents (nodes of the graph) can provide extra information compared to the initial patent query for finding and recommending citations. Contextual similarity between a pair of documents is defined as having common classification codes, common inventor or common applicant.

We performed this analysis on the relevance judgments of CLEF-IP 2010. Figure 1 shows the results of this analysis per topic. Please note that we randomly selected a subset of 100 topics in order to highlight the differences between the two sets. We plotted the number of relevant documents for each topic that have common attributes with the query document. We also plotted the number of relevant documents that have similar attributes with the documents in the citation-graph. We extracted the information about the

[2] http://www.ifs.tuwien.ac.at/~clef-ip/download-central.shtml

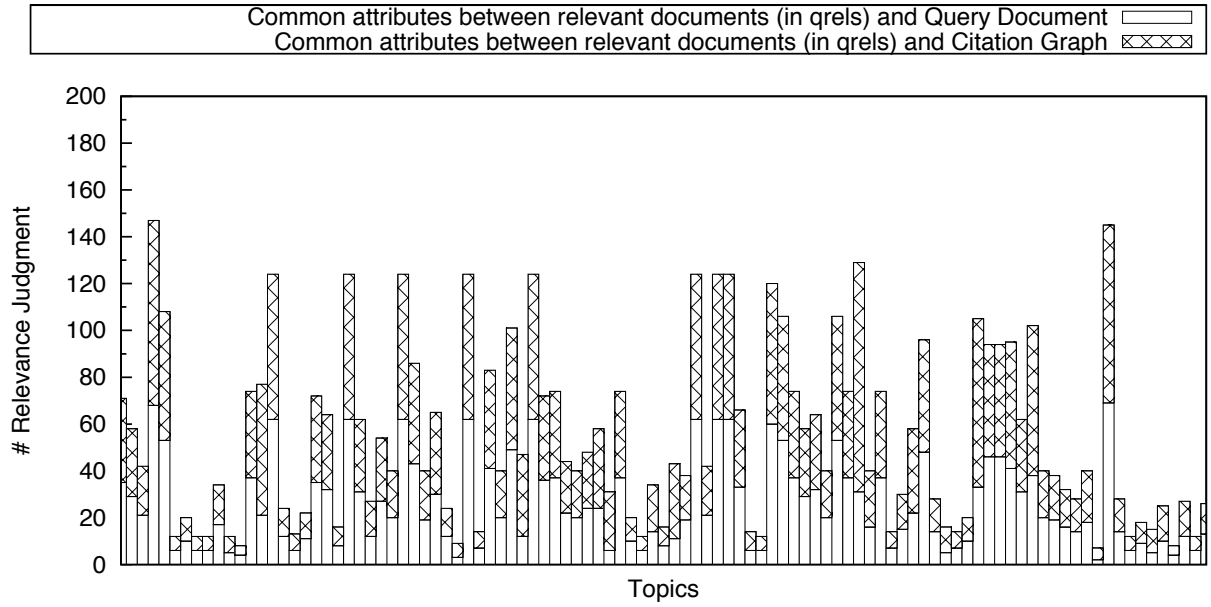

Figure 1: Contextual attribute similarity between relevant documents (the ground truth) and the query patent vs. the contextual attribute similarity between relevant documents (the ground truth) and the documents in the topic-specific citation graph.

relevant documents associated to each topic from the relevance judgments. From this histogram we can observe that for almost all the topics, contextual attribute similarity between the relevant documents and the topic-specific citation graph is higher than the contextual attribute similarity between the relevant documents and the query.

This analysis shows that the contextual attributes extracted from the citation network provide additional and complementary information compared to the initial query patent. This investigation answered our first research question regarding the usefulness of employing citation network structure for distinguishing between relevant and non-relevant documents.

In this paper we propose a method that uses this additional information to better distinguish between relevant and non-relevant documents and increase the accuracy of recommending citation for a given patent application.

3.4 Building the Network

In the CLEF-IP dataset, the citations of patent queries have been removed by the organizers and used for building the relevance judgments. However, we have access to the citations of all other documents apart from the patent queries in the collection. We used a web service offered by the European Patent Office (EPO)[3] to extract all the citations of the documents in the collection except the query documents.

As previous work [5] suggests, computing Page Rank values as a measure of static document quality (calculated independently of any query a system might receive) has a clear disadvantage compared to conditioning the computation of Page Rank values on the query being served. Thus we will focus on how to assemble a subset of patent documents around the topic of the query, from the graph induced by their citation links. By doing so we are able to derive Page Rank values relative to particular queries. Our approach is inspired by the HITS algorithm [7] where a small sub graph of the entire web related to the query (as opposed to the whole web) is chosen for estimating the importance of a webpage.

We build such a graph by gathering a subset of linked documents in the patent collection related to a query following the two steps below:

1. Given a query patent, we perform a search and we retrieve an initial ranked list of documents. We take the top-k documents from this primary rank list and call this the *root set* of documents.
2. We construct the *base set* of documents, by including the root set as well as any document that either cites a document in the root set, or is cited by a document in the root set.

In our citation network, each node denotes a patent document in the graph and each link denotes a citation relationship between two patent documents. The edges are weighted according to the similarity between the target and source nodes. The similarity metrics denoting the edge weights are defined in the next section. The linking edges among the nodes reveal a lot of valuable information about the potential relevance propagation over the network.

4. INITIAL QUERY MODEL (BASELINE)

We first focus on the textual content available in the initial query and build a query model as explained below. We estimate a bigram language model using SRILM [17], trained

[3] `http://www.epo.org/searching/free/ops.html`

on the patent documents. For this we used the Katz backoff smoothing [6] method. Accordingly, we estimated a query by quantifying the difference between the language model of the query q and the language model of the collection. We calculated the cross entropy between q and the language model of the collection LM_{col} as follows:

$$H(q, LM_{col}) = -\frac{1}{N}\sum_{i} logP_{LM_{col}}(b_i) \quad (1)$$

where $b_1, ..., b_N$ are the word bigrams found in q and $P_{LM_{col}}(b_i)$ is the probability of the bigram b_i under the language model of the collection. Higher cross-entropy values indicate bigrams that are distinguishing the language usage of the query patent from the language usage of the collection. We refer to this query model as Q_{Orig} in the rest of the paper.

5. TIME-AWARE NETWORK ANALYSIS

In this section we explain the time-aware Page Rank analysis we performed. We first discuss how the initial probability of selecting a node is discounted based on the age of the node to address the bias introduced by Page Rank against recent documents [1]. We then describe how similarity functions are used to associate strength/weight to edges and how this affects the transition probabilities. Finally we describe the calculation of weighted Page Rank scores for the nodes in the network and how these scores are used as a document prior to guide the term extraction process.

Initial Probability. We modify the initial probability to assign importance to newer nodes as described in Equation 2. The initial probability distribution is exponentially discounted according to the age of the nodes. This takes into account freshness of documents and it is in contrast with the original Page Rank seeds where a uniform distribution is assumed over all nodes.

$$\rho_i = e^{\frac{-age}{\tau_d}} \quad (2)$$

where ρ_i is the initial probability for selecting a node that discounts the node according to its age. τ_d is a time granularity which could be set to a month or a year. Note that we use the publication date of the first filed patent as the time tag to calculate its age. The unit of time granularity considered in our analysis is a year.

Transition Probability Matrix. The transition probability matrix W is described as follows:

$$W_{ij} = \begin{cases} \frac{w_{ij}}{\Sigma_k w_{ik}} & \text{if } i \text{ cites } j; \\ 0 & \text{otherwise;} \end{cases} \quad (3)$$

where w_{ij} defines the weight of the edge between node i and node j. Edge weights are normalized, therefore, the sum of weights for all outgoing edges from each vertex equals 1. Each entry w_{ij} defines the conditional probability that a walk will traverse edge (i, j) given that it is currently at node i.

We consider different similarity functions to assign weights to edges based on textual and contextual similarities between two nodes. We use different similarity metrics such as common classification codes, common inventor and common applicant to assign weights to edges in the citation graph. Table 1 lists different similarity metrics we used.

Weights for citation relationship (v_i, v_j)	
Weight	Description
SIM_{IPC}	Number of common classification codes
SIM_{LEX}	Lexical similarity (cosine similarity)
SIM_{TEMP}	Difference between the publication dates of v_i and v_j (in years)
$SIM_{Inventor}$	Number of common inventors
$SIM_{Applicant}$	Number of common applicants

Table 1: Similarity metrics denoting edge weights on the citation graph. Each metric is defined between the source node and the target node.

We construct a separate transition probability matrix (as defined in Equation 3) according to each edge type mentioned in Table 1. For instance, $W(SIM_{IPC})$ denotes the transition probability matrix in which the edge weight w_{ij} is proportional to the number of similar classification codes between nodes i and j.

Weight Combination. Our goal is to see if the combination of different similarity functions for calculating the edge weights can improve the result of a single similarity function. To this end, we combine different edge weights (mentioned in Table 1) between node i and j as shown in Equation 4. We adopt from [21] the following function for combining the weights of each w_{ij} entry.

$$p_{ij} = \begin{cases} 1 - \exp(L) & \text{if } 0 \leq L \leq 0.5; \\ 1 - \exp(-L) & \text{if } 0.5 < L < 1; \end{cases} \quad (4)$$

$$L = \prod_{w_{ij} \in W(SIM)} w_{ij}$$

$$a_{ij} = \begin{cases} 1 & \text{if } p_{ij} \geq 0.5; \\ 0 & \text{if } p_{ij} < 0.5; \end{cases} \quad (5)$$

a_{ij} is an entry of the adjacency matrix after weight combination. We refer to this weight combination method as SIM_{COMB}. We will compare different similarity functions in Section 7.2.

Page Rank. The Page Rank formula is denoted as:

$$PR(u) = \frac{\lambda}{N} + (1 - \lambda) \times \Sigma_{v \in B_u} \frac{PR(v)}{L_v} \quad (6)$$

where N is the number of nodes under consideration, $B(u)$ is the set of nodes that point to u, and L_v is the number of outgoing links from node v [3]. We consider a weighted version of the Page Rank formula as follows:

$$PR{-}W(u) = \lambda \times \frac{O_u}{\Sigma_{p \in L(v)} O_p} + (1{-}\lambda) \times \Sigma_{v \in B_u} \frac{PR(v)}{L_v} \quad (7)$$

where O_u and O_p represent the number of outgoing edges of nodes u and p. We decided to use the weighted Page Rank as we observed a better performance compared to the conventional Page Rank in terms of the retrieval effectiveness of the final ranked list. We will explain these results in Section 7.2. We now need to answer the question of how to formulate a query given the edge weights and their combination.

6. RE-RANKING AND QUERY EXPANSION

Our approach for query modeling aims to improve the language model of the initial query by using the term distribution of documents in the citation network. We use document metadata (node attributes) and network structure to find important documents in the citation network. The key assumption of this paper is that the term distribution of documents with more importance in the citation graph will help to overcome the term mismatch problem and thus term selection from them is more effective.

We identify and weigh the most distinguishing terms in the documents in the citation graph and then use the calculated weighted Page Rank value as a document prior in a language modeling framework. This term sampling is performed as follows (the terms are ranked by the following score):

$$P(t|Q_{cit}) = Z_t \sum_{D \in G_{cit}} P(t|D)P(D) \quad (8)$$

where G_{cit} is the citation graph. $P(D)$, the weight of each document, is proportional to its Page Rank score calculated according to Equation 7. Z_t is a normalization factor.

We interpolate the citation query model with the initial query model (as estimated in Equation 1):

$$P(t|Q) = \lambda' \cdot P(t|Q_{orig}) + (1 - \lambda') \cdot P(t|Q_{cit}) \quad (9)$$

The M highest terms from the updated query model are then used as a query to retrieve a final ranked list of documents.

7. EXPERIMENTAL EVALUATION

Here we describe the details of our experimental setup, reporting about the test sets and baselines. We report the results of our experiments with the initial estimated query model from the patent document. We then show how the expanded query model estimated from the documents in the patent citation network improves over the initial query model.

7.1 Experimental Setups

In the following we first describe the datasets used in our experiments and provide some statistics about them. We then proceed to discuss some of the details of the methodology used for our analysis. In particular, here we focus on the fields extracted from a patent document.

Patent Dataset. In this study, we used two large patent data collections released by the Intellectual Property track at CLEF, called CLEF-IP 2010 and CLEF-IP 2011. CLEF-IP 2010 consists of 1.3 million distinct patent documents published between 1985 and 2001, from which we extracted 504,110 companies and 2,711,471 inventors. CLEF-IP 2011 consists of about 1.3 million distinct patent documents from which we extracted 600,001 companies and 2,712,298 inventors. There are 208 nodes and 1066 edges on average in each topic-specific citation graph built from the dataset. These statistics correspond to a case where 20 feedback documents are used as the root set for building the citation network.

In our experiments we used the English subsection of both collections. The English test set of CLEF-IP 2010 corresponds to 1348 topics (patent applications). The English test set of CLEF-IP 2011 corresponds to 1351 topics. We used the training topics of CLEF-IP 2010 for tuning the parameters of our model. This training set consists of 300 topics. We used five-fold cross validation.

Preprocessing. We performed stemming using porter stemmer and removed stop words according to the Terrier[4] general stop word list. We also performed token and sentence segmentation on the documents.

Evaluation. We quantify the performance of our proposed method by reporting Mean Average Precision (MAP) and recall. We also report the evaluation results of our approach in terms of Patent Retrieval Evaluation Score (PRES) [11] which is specifically designed for recall-oriented applications. PRES metric is calculated as follows:

$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{\max}}$$

$$\sum r_i = \sum_{i=1}^{nR} r_i + nR(N_{\max} + n) - \frac{nR(nR-1)}{2}$$

where $N_{\max}$ is the number of documents to be checked by the user (cut-off value), n is the number of relevant documents, and $\sum r_i$ is the summation of ranks of relevant documents.

We used the Language Modeling approach with Dirichlet smoothing [23] to score documents and build the initial rank list. We empirically set the value for the smoothing parameter to 1500. We also used Language Modeling for the re-ranking of the results. We select top k documents from the primary rank list to generate the candidate documents for suggesting as citations. The number of k is empirically set to 20 in our experiments. We will study how this parameter affects the recommending accuracy and coverage in Section 7.2. The reported results for the methods are obtained using 100 terms with highest weights selected from the estimated expanded query model. We experimented with different number of query terms and chose the best value.

Baselines. Table 3 reports the performance of the top ranked participants in CLEF-IP 2010. The best performing run [8] (labeled as BAS1) uses initial citations provided by the patent applicant, which are extracted by training a Conditional Random Field (CRF). BAS1 method employs two complementary indices, one constructed by extracting terms from the patent collection and the other built from external resources such as Wikipedia. The second best run [10] (labeled as BAS2) formulates a query from the query patent application by extracting its most frequent unigrams and bigrams.

method	MAP	recall	PRES
BAS1 [8]	0.226	0.6946	0.615
BAS2 [10]	0.136	0.5886	0.483

Table 3: Evaluation results of the best participating teams in CLEF-IP 2010 challenge.

As implied before, the relevance judgments for the CLEF-IP challenge are built based on the report provided by the

[4] http://terrier.org/

Method	metric	SIM_{IPC}	SIM_{LEX}	SIM_{TEMP}	$SIM_{Inventor}$	$SIM_{Applicant}$	SIM_{COMB}
AQE-PR (uniform PR seed)	MAP@1000	0.0794	0.1073	0.1183	0.1308	0.0776	0.07800
	recall@1000	0.6705 †	0.6657 †	0.6170	0.6186	0.5978	0.6768 †
	PRES@1000	0.5386 †	0.5370 †	0.4930	0.5034	0.4600	0.5784 †
AQE-TPR (age-based PR seed)	MAP@1000	0.1351	0.1328	0.1306	0.1560 †	0.1025	0.1468 †
	recall@1000	0.6662 †	0.6614 †	0.6156	0.6107	0.5942	0.6750 †
	PRES@1000	0.5676 †	0.5621 †	0.5004	0.5040	0.4809	0.5850 †

Table 2: Performance comparison of query models built after computing random walks on the citation graph using different edge weights. The symbol † denotes statistical significant improvement over BQE. Wilcoxon signed ranked matched pairs test with a confidence level of 0.01 was used for testing statistical significance.

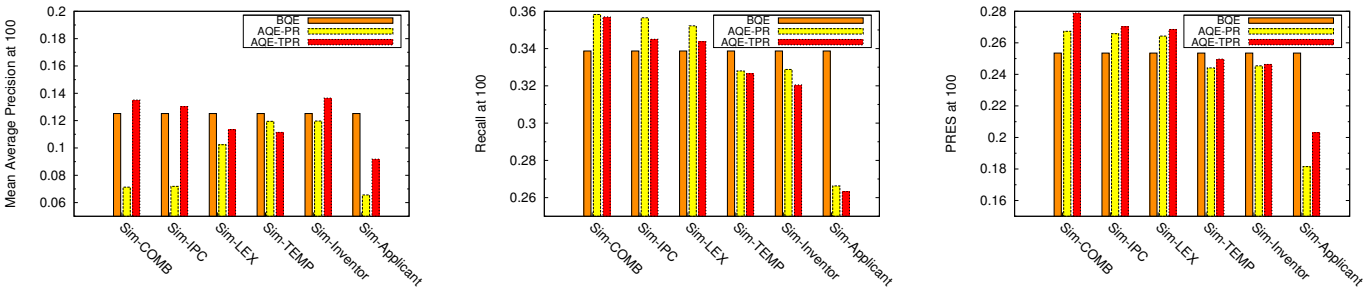

Figure 2: Performance results of BQE, AQE-PR and AQE-TPR using different similarity functions. (a) MAP@100. (b) Recall@100. (c) PRES@100.

patent examiner. This report can have overlapping references with the initial citation list provided by the patent applicant. Thus, using the initial citation list provided by the patent applicant, as performed by BAS2, raises a concern about the validity of the evaluation of the task. To have a fair comparison setting, we can not compare our work directly to the first approach presented in Table 3. However, the second approach shows the performance of a method which does not take into account the initial citations provided by the patent applicant. Thus, we can compare our method directly to the second best approach. Note that our proposed approach does not take into account the initial citations provided by the patent applicant.

Table 4 shows the performance of the primary rank list before query expansion (labeled as BQE), built from the patent application topic (as implied in Section 4). These results are comparable to the BAS2. We will use the results of BQE method as the baseline in the rest of this paper.

Query model built from query patent application			
method	MAP@1000	recall@1000	PRES@1000
BQE	0.1363	0.6231	0.5076

Table 4: Performance of the primary rank list (before query expansion) on CLEF-IP 2010 .

7.2 Experiments

In this section we describe the experiments that we conducted to evaluate the usefulness of our proposed method and present the results.

We first describe the structure of our experiments. Our goal is to see if the combination of different similarity functions for calculating the edge weights can improve the result of a single similarity function. We also compare the effectiveness of the query expansion method with uniform initial Page Rank seeds (labeled as AQE-PR) with the proposed query expansion method using the temporal Page Rank seeds (labeled as AQE-TPR). We compare the performance of these methods to the baseline (labeled as BQE). We then study the effect of the number of feedback documents considered while generating the query-specific citation network in Section 7.2.1. We further investigate the effect of the estimated query models and present the evaluation results of AQE-PR and AQE-TPR on different technological fields inferred from IPC classes in Section 7.2.2. We show some example queries and the list of terms selected from them in Section 7.2.3. Finally, we report the evaluation results of our approach on CLEF-IP 2011 in Section 7.2.4.

We now investigate the effect of using different similarity functions for calculating edge weights on the final performance of our method. These results are presented in Table 2. In Table 2, we can notice an interesting effect in the results of SIM_{COMB} by comparing AQE-PR to AQE-TPR: the MAP results of SIM_{COMB} improved enormously without impacting the recall value. This can be attributed to the effect of the initial age-based seed values used for random walks. We can see a similar trend of MAP improvement for all the similarity metrics while maintaining a steady recall value. This improvement highlights the importance of using the age-based penalty for the random walker.

In addition to this, looking at the data in Table 2, we see that similarity $Sim_{Applicant}$, $Sim_{Inventor}$ and Sim_{Temp} have curiously lower recall values with respect to Sim_{COMB}, Sim_{LEX} and Sim_{IPC}. The result of this study suggests that the contextual attributes like applicant, inventor and temporal

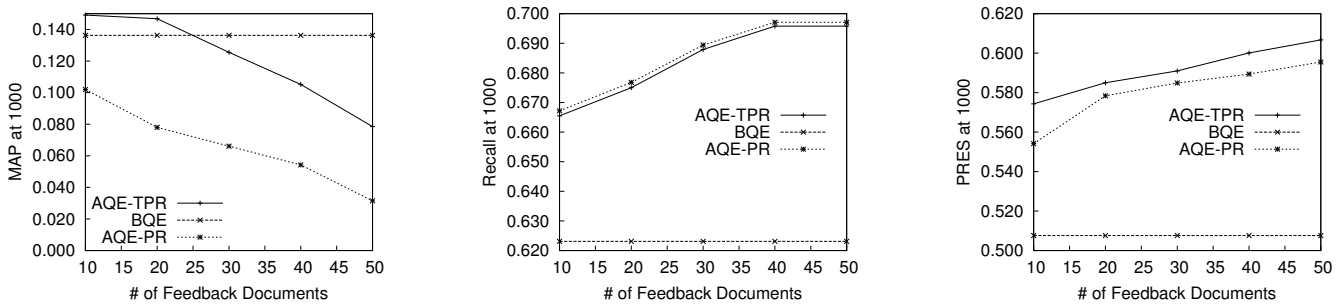

Figure 3: The effect of varying the number of feedback documents on AQE-TPR(SIM_{COMB}), AQE-PR(SIM_{COMB}), and BQE. (a) MAP@1000. (b) Recall@1000. (c) PRES@1000.

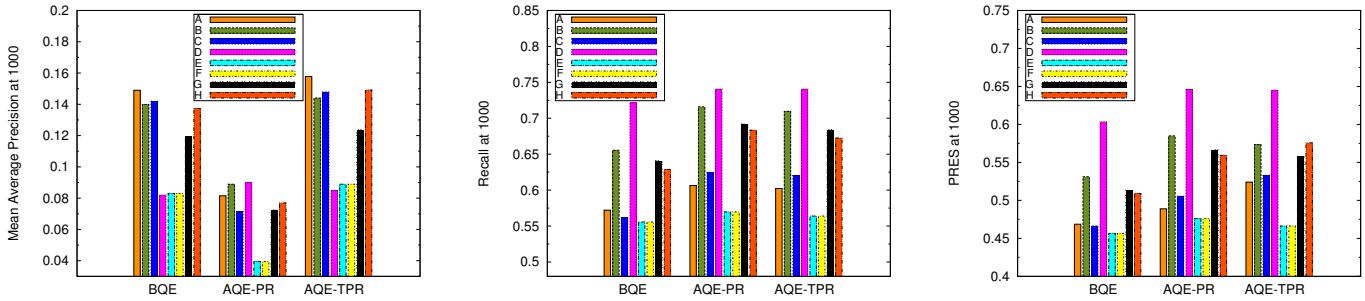

Figure 4: Evaluation results over CLEF-IP test set queries grouped according to technological categories. (a) MAP@1000. (b) Recall@1000. (c) PRES@1000.

information can be helpful when used jointly with lexical (content) attributes and topical aspects (denoted by IPC) but not separately. In fact, by excluding the results of the combination Sim_{COMB}, we can see that the best recall value is achieved through the use of topical aspects. The second best result is achieved by content attributes. The symbol † in Table 2 denotes statistical significant improvement over BQE. We used Wilcoxon signed ranked matched pairs test with a confidence level of 0.01 level for testing statistical significance. These findings provided answer to our third research question.

We also compare the performance of our method with and without taking into account the temporal properties of documents in the modeling process. The result presented in Table 2 shows how integrating the temporal properties in AQE-TPR improves MAP and PRES values over AQE-TPR which does not use temporal properties. Figure 2 shows the MAP, recall and PRES values at cut-off rank 100. This analysis confirms that our model works well when combining all the attributes together at different cut-off values. We also considered the linear combination of the weights but the results were not satisfactory. Thus we did not present these results in the experimental section.

7.2.1 *Effect of Number of Feedback Documents*

As AQE-PR and AQE-TPR share the parameter k (i.e. number of feedback documents used as the root set for building the citation network) we are interested in understanding how this parameter affects the retrieval performance of these two methods. We thus draw the sensitivity curves of AQE-PR and AQE-TPR with respect to k in Figure 3.

There is a marked drop of performance in terms of MAP when we increase the number of feedback documents. However, increasing the number of feedback documents has a positive effect in terms of recall and PRES. The results of Figure 3 show that AQE-TPR is consistently more effective than AQE-PR when we vary the value of k. In fact, we can see that the curve trends of AQE-PR and AQE-TPR are similar to each other and they share the same optimal setting of k. The reported results for AQE-PR and AQE-TPR elsewhere in this paper are obtained using the top 20 feedback documents to generate the root set, since it has a good balance between MAP and recall. Top 100 terms are selected from the expanded query model.

7.2.2 *Results on Separate IPC Classes*

International Patent Classification (IPC)[5] divides technology domains into eight different fields. Some information about the field of technology of the query topics in CLEF-IP 2010 dataset are presented in Table 5.

We present the evaluation results of our methods in different technological fields in Figure 4. The results of Figure 4 demonstrate that AQE-TPR outperforms BQE consistently in terms of recall over all topical aspects (IPC classes) and also achieves better MAP and PRES scores than BQE in most cases.

[5]See `http://www.wipo.int/classifications/ipc/en/` for such classes

Category	Description	# of topics
A	Human Necessities	154
B	Performing Operations and Transporting	307
C	Chemistry and Metallurgy	255
D	Textiles and Papers	10
E	Fixed Constructions	7
F	Mechanical Engineering, Heating, Weapons, and Blasting	90
G	Physics	289
H	Electricity	236
	total number of queries	1348

Table 5: IPC section distribution over English test set of CLEF-IP 2010.

Topic Number	Query Patent Document (BQE)	Citation Graph using AQE-TPR (SIM_{COMB})
PAC-9	receiver, wireless microphone, transmit, operator, communication, display	signal level, lan wireless, mobile network bluetooth, transact, station, broadcast
PAC-999	lithographic printing, hydrophilic surface, thermosensitive, polymer, image	printing plates, print precursor, pigment, heat laser, ink, water soluble, dissolution
PAC-997	electron beams, deflection, cathode, ray, electron gun, shadow	ray tube, funnel neck, panel, emit deflection yoke, magnet, coil, rectangular

Table 6: Top keywords, phrases extracted from example topics using the BQE and AQE-TPR methods.

Topic Number	PAC-999
Patent Application Number	EP-1356929-A2
Title	Method of preparation of lithographic printing plates
Abstract	A method for preparation of a lithographic printing plate, which comprises the steps of: imagewise recording on a lithographic printing plate precursor comprising a support having a hydrophilic surface and a thermosensitive layer, the thermosensitive layer comprising at least one of polymer particles and a microcapsule encapsulating an oleophilic compound therein; and rubbing the printing plate precursor by a rubbing member in the presence of a processing liquid to remove the thermosensitive layer of non-image portions.

Table 7: An example topic in CLEF-IP 2010 test set.

Regarding different technological categories derived from IPC classes, we observe that MAP improvements on A, E and H categories are as high as 6%-8% for AQE-TPR method over the baseline BQE. Also, the recall improvements on B, C and F categories are as high as 7%-10% for AQE-TPR method over the BQE. The results of Figure 4 show that AQE-TPR obtains better PRES values compared to BQE especially in categories A, C and H (up to 0.11%-0.14%). These are interesting observations, which suggest that using the contextual information derived from the citation network achieves the best improvement in topics related to Chemistry, Metallurgy and Electricity. We leave further investigation to the future work.

7.2.3 Example Extracted Topic Keywords

We list the extracted keywords (unigrams and bigrams) for some example topics in Table 6. For topic PAC-9 with the title "Wireless Microphone Communication System", we can see that the keywords extracted from the query patent document and the keywords extracted from the citation graph differ from each other. This shows that the keywords extracted for the same topic but from two different sources are adequately discriminative. To familiarize the reader with different fields in a patent application, an example is shown in Table 7.

7.2.4 Comparison on CLEF-IP 2011 dataset

Table 8 reports the evaluation results of the best participating teams in CLEF-IP 2011 challenge [16], the results of BQE method, together with the best results we obtained using the AQE-TPR. Note that the PRES metric was not reported in the official results. It can be seen from Table 8

method	MAP	recall	PRES
nijm (rank 1)	0.0582	0.6303	NA
hyder (rank 2)	0.0593	0.5713	NA
BQE	0.0990	0.5935	0.4859
AQE-TPR ($SIM_{Applicant}$)	0.0771	0.5810	0.4910
AQE-TPR ($SIM_{Inventor}$)	0.1365 †	0.5887	0.5104
AQE-TPR (SIM_{TEMP})	0.1090	0.5934	0.5120
AQE-TPR (SIM_{LEX})	0.1135 †	0.6280 †	0.5276
AQE-TPR (SIM_{IPC})	0.1198 †	0.6351 †	0.5305 †
AQE-TPR (SIM_{COMB})	0.1250 †	0.6470 †	0.5363 †

Table 8: Performance comparison of AQE-TPR with other approaches on CLEF-IP 2011 dataset at (cut-off 1000). The symbol † denotes statistical significant improvement over BQE. Wilcoxon signed ranked matched pairs test with a confidence level of 0.01 was used for testing statistical significance.

that the AQE-TPR (SIM_{COMB}) method achieves better results compared to other approaches in terms of recall, MAP, and PRES. The improvements achieved by AQE-TPR (SIM_{COMB}) are statistically significant in comparison to BQE method. We answered our second research question based on the evaluations performed over CLEF-IP datasets presented in Tables 2 and 8.

8. CONCLUSION AND FUTURE WORK

In this paper we study the problem of finding and recommending patent citations for a given query patent (patent application). We extracted different types of metadata from the query patent and the dataset such as classification information (implying topical aspects), applicant names, inventor names and publication dates. We built a directed weighted graph of patent citations and developed a framework that combines network structure and node attributes to infer and discover similarities among patent documents. We performed this by estimating a query model which employs the network structure and node attributes. We evaluated our proposed model on the CLEF-IP datasets and the experimental results showed that our model achieved significant improvements in terms of recall and MAP over competitive state of the art prior art search approaches.

As for the future work, it would be interesting to model the graph as a heterogeneous network. Our current model calculates the strength for the edges on the network by looking into different similarity metrics utilizing textual and contextual attributes at node level. However, considering the attributes directly as nodes on the network might lead to a better diffusion schema of information on the network.

9. ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their valuable feedback that helped us improve the quality of this paper.

10. REFERENCES

[1] R. A. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web structure, dynamics and page quality. In *Proceedings of String Processing and Information Retrieval (SPIRE)*, pages 117–130, 2002.

[2] S. Bashir and A. Rauber. Improving retrievability of patents in prior-art search. In *Proceedings of ECIR*, pages 457–470, 2010.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

[4] E. D'hondt, S. Verberne, C. H. A. Koster, and L. Boves. Text representations for patent classification. *Computational Linguistics*, 39(3):755–775, 2013.

[5] A. Fujii. Enhancing patent retrieval by citation analysis. In *Proceedings of SIGIR*, pages 793–794, 2007.

[6] S. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, 1987.

[7] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[8] P. Lopez and L. Romary. Experiments with citation mining and key-term extraction for prior art search. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[9] M. Lupu and A. Hanbury. *Patent Retrieval*. Foundations and Trends in Information Retrieval, 2013.

[10] W. Magdy and G. J. F. Jones. Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[11] W. Magdy and G. J. F. Jones. PRES: A score metric for evaluating recall-oriented information retrieval applications. In *Proceedings of SIGIR*, pages 611–618, 2010.

[12] P. Mahdabi, L. Andersson, M. Keikha, and F. Crestani. Automatic refinement of patent queries using concept importance predictors. In *Proceedings of SIGIR*, pages 505–514, 2012.

[13] P. Mahdabi, S. Gerani, J. X. Huang, and F. Crestani. Leveraging conceptual lexicon: Query disambiguation using proximity information for patent retrieval. In *Proceedings of SIGIR*, pages 113–122, 2013.

[14] D. M. Mimno and A. McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of KDD*, pages 500–509, 2007.

[15] S. Oh, Z. Lei, W.-C. Lee, P. Mitra, and J. Yen. CV-PCR: a context-guided value-driven framework for patent citation recommendation. In *Proceedings of CIKM*, pages 2291–2296, 2013.

[16] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz:. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.

[17] A. Stolcke. SRILM – an extensible language modeling toolkit. *In Proceedings of ICSLP*, pages 901–904, 2002.

[18] J. Tang, B. Wang, Y. Yang, P. Hu, Y. Zhao, X. Yan, B. Gao, M. Huang, P. Xu, W. Li, and A. K. Usadi. PatentMiner: topic-driven patent analysis and mining. In *Proceedings of KDD*, pages 1366–1374, 2012.

[19] W. Tang, J. Tang, T. Lei, C. Tan, B. Gao, and T. Li. On optimization of expertise matching with various constraints. *Neurocomputing*, 76(1):71–83, 2012.

[20] S. Wu, J. Sun, and J. Tang. Patent partner recommendation in enterprise social networks. In *Proceedings of WSDM*, pages 43–52, 2013.

[21] J. Yang and J. Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of WSDM*, pages 587–596, 2013.

[22] Y. Yang, J. Tang, J. Keomany, Y. Zhao, J. Li, Y. Ding, T. Li, and L. Wang. Mining competitive relationships by learning across heterogeneous networks. In *Proceedings of CIKM*, pages 1432–1441, 2012.

[23] C. Zhai and J. D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of SIGIR*, pages 334–342, 2001.