Link Prediction in Citation Networks

Naoki Shibata and Yuya Kajikawa
Innovation Policy Research Center, Graduate School of Engineering, The University of Tokyo, 2-11-16 Yayoi, Bunkyo, Tokyo, 113-8656, JAPAN. E-mail: {shibata, kaji}@ipr-ctr.u-tokyo.ac.jp

Ichiro Sakata
Innovation Policy Research Center, Graduate School of Engineering, The University of Tokyo, 2-11-16 Yayoi, Bunkyo, Tokyo, 113-8656, JAPAN and Todai Policy Alternatives Research Institute, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo, 113-0033, JAPAN. E-mail: isakata@ipr-ctr.u-tokyo.ac.jp

In this article, we build models to predict the existence of citations among papers by formulating link prediction for 5 large-scale datasets of citation networks. The supervised machine-learning model is applied with 11 features. As a result, our learner performs very well, with the F1 values of between 0.74 and 0.82. Three features in particular, *link-based Jaccard coefficient*, *difference in betweenness centrality*, and *cosine similarity of term frequency–inverse document frequency vectors,* largely affect the predictions of citations. The results also indicate that different models are required for different types of research areas—research fields with a single issue or research fields with multiple issues. In the case of research fields with multiple issues, there are barriers among research fields because our results indicate that papers tend to be cited in each research field locally. Therefore, one must consider the typology of targeted research areas when building models for link prediction in citation networks.

Introduction

Most phenomena in the real world are complex and dynamic; new nodes and links are added so frequently that the structures of the entire network change and grow quickly over time. Therefore, understanding the fundamental growth mechanism of social networks is still difficult but significant.

Because the number of academic papers exponentially increases (Price, 1965), each academic area becomes specialized and segmented. This is because the individual scientist has to focus on or specialize in only a few scientific subareas to keep up with the growth of the areas as the number of academic papers increases. Davidson, Hendrickson, Johnson, Meyers, and Wylie (1998, p. 259) describe this situation as follows: "For most of history, mankind has suffered from a shortage of information. Now, in just the infancy of the electronic age, we have begun to suffer from information excess." Recent researches of science mapping (Small, 2003; Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2009), clustering and visualization (Chen, 1999; Small, 1999; Börner et al., 2003), and emerging topic detection (Boyack, Wylie, & Davidson, 2002; Chen, Cribbin, Macredie, & Morar, 2002; Shibata, Kajikawa, Takeda, & Matsushima, 2008) have pointed out gaps among traditional research fields. Under this circumstance, the individual scientist has to focus on or specialize in only a few scientific subdomains to keep up with the growth of the domains, which means that researchers must focus on increasingly narrowing domains. Therefore, segmentation occurs simultaneously with specialization, which gives rise to severe problems but also the opportunity to find crucial knowledge by integrating different domains (Kajikawa, Abe, & Noda, 2006).

This article uses the following research question to investigate the mechanism of why papers are cited: What factors affect the existence of links using features intrinsic to the network itself, namely, *link prediction*, which will help scholars to know which paper to cite and managers to identify future core papers? There are a number of motivations and factors for a paper to cite or be cited (MacRoberts & MacRoberts, 1989). In this article, we utilize textual, topological, and attribute features for link prediction, which are considered to influence citing behaviours. Consider a citation network among academic papers, for example. There are many reasons, exogenous to the network, why author(s) cite another paper written by other author(s) who have never worked together. For instance, two papers that cite (or are cited by) several papers in common are more likely to cite (or to be cited), which can be modeled by topological features of the

Received May 10, 2011; revised August 13, 2011; accepted August 15, 2011

 • Published online 13 October 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21664

citation network. Another example is two papers dealing with similar topics are likely to cite each other, which can be modeled by textual features. As such, a large number of new citations are hinted at by the topology and textual trait of the citation network; two papers that are in "close" or "similar" positions in the citation network will cite or be cited.

Several link prediction models have been proposed. Liben-Nowell and Kleinberg (2003) proposed a model for link prediction in large coauthorship networks. Clauset, Moore, and Newman (2008) investigated the hierarchical structure of social networks to predict missing connections in partially known networks with high accuracy. Popescul and Ungar (2003) proposed a new approach for Statistical Relational Learning to build link prediction models. Hasan, Chaoji, Salem, and Zaki (2006) tested several supervised learning models (decision tree, k-nearest neighbor, multilayer perception, support vector machine [SVM], radial basis function [RBF] network) for link predictions and concluded that SVM outperformed all of them by a narrow margin in all performance measures. Murata and Moriyasu (2008) applied the model of Liben-Nowell and Kleinberg to social networks of Question-Answering Bulletin Boards. Caragea, Bahirwani, Aljandal, and Hsu (2009) proposed an algorithm to predict potential friendships based on a clustering approach in *LiveJournal*, a social network journal service with a focus on user interactions. Lu, Jin, and Zhou (2009) presented a local path index to estimate the likelihood of the existence of a link between two nodes. Seglen (1994) analysed the trends of papers in the journals with large impact factors. Vinkler and Davidson (2002) indicated that the papers in growing journals in terms of the number of papers are more likely to be cited. Hwang, Wylie, Wei, and Liao (2010) proposed recommendation engines based on the coauthorship networks.

However, there is still room to improve link prediction models optimized for citation networks. As described above, several link prediction models with different learning approaches and different features have been proposed for different social networks. First, our focus is on citation networks. Second, we apply SVMs as our supervised learning method, as SVM is the best learner according to Hasan et al. (2006). Third, we use more comprehensive features optimized for citation networks. We use 11 features, which can be categorized into three types; topological, semantic, and attribute features. Topological features are calculated by the topological positions (in citation networks) of two nodes in a given pair, while semantic features are calculated by the semantic similarity between two documents in a given pair. Attribute features are relational values of a pair.

In this article, we apply supervised machine learning models to predict the existence of citations with five large-scale citation datasets. Because the number of papers has increased and research areas have been segmented, it has been more difficult for both researchers and reviewers to decide which papers should be cited. As a kind of socially intelligent computing, modeling the underlying factors, which affects the existence and the class of links, helps us make decisions whether to link more accurately even with a huge number of data. One possible application will be a citation recommendation system for authors of scientific publications and patents. This system recommends possible candidate papers to be cited based on the algorithm, given the attributes of the documents. With this system, authors can retrieve the documents they should cite in their papers. The reviewers of scientific papers can reduce their time to check whether the references in those papers are adequate or not. Second, well-organized link prediction can reveal how and why authors cite other scientific papers. Finally, link prediction can bond different research fields with similar topics but from different disciplines.

In the next section, we describe the entire procedure of this article. In the following section, we illustrate the results of the training and testing. Then we discuss the results to predict the existence of citations, followed by conclusions.

Research Methods

In this section, the methodology of link prediction, how the features are selected, and the experiment procedure are described.

Methodology of Link Prediction

We predict the existence of each citation given a citation network. As the link prediction problem, we predict whether there exists a link, a citation from a certain paper X_i to another paper X_j given a pair (X_i, X_j) on a citation network. More formally, let $(X_i, X_j) := \{x_{ij1}, \ldots, x_{ijs}\}$ denote the s attributes where each x_{ij} represents the attribute regarding a link between node X_i and node X_j in the given citation network. Then, let $Y := \{y\}$ denote the existence of a citation, where y takes a value either of $+1$, which indicates a "citation exists" or -1 (which indicates a "citation does not exist").

Our goal is to build a hypothesis h that relates the paper pair instances of (X_i, X_j) to the existence of citations, i.e., $Y = h(X_i, X_j)$. Several methods of building such a model have been developed in the field of machine learning. In our study, we employ the SVM, which is a state-of-the-art predictive model (Vapnik, 2000). The SVM is well-known for its high predictive performance and has been applied in numerous application areas. The SVM assumes the following linear model:

$$y := signh(x) := sign(w^T x) := sign(w_1 x_1 + w_2 x_2 + \cdots + w_d x_d),$$

where $x = (x_1, x_2, \ldots, x_d)$ is a d-dimensional feature vector and $w = (w_1, w_2, \ldots, w_d)$ is the parameter vector of the same dimension that specifies the model. A positive value of w_j indicates that the j-th feature x_j positively contributes to the prediction, while a negative value contributes to it negatively. The sign function returns $+1$ when its argument is positive, and returns -1 otherwise. Given the data set X and Y, the SVM learning algorithm finds the optimal parameter $w*$ that minimizes the following objective function:

$$\sum_i \max\{1 - y_i h(x_i), 0\} + c\|w\|_2^2,$$

DOI: 10.1002/asi

TABLE 1. Features of supervised machine learning model.

Feature	Type of value
1. No. common neighbors	Integer
2. Link-based Jaccard coefficient	Float [0,1]
3. Difference in betweenness centrality	Float
4. Difference in the number of in-links	Float
5. Is same cluster	Binary
6. Cosine similarity of tf-idf vectors	Float [0,1]
7. Difference in publication year	Integer
8. The number of common authors	Integer
9. Is self-citation	Binary
10. Is published in same journal	Binary
11. Number of times "to" cited	Integer

where the first term is the loss function, which penalizes misclassifications, and the second term is the regularization term, which avoids over-fitting to the given data set. c is a small constant that balances the two terms. The regularization term $||w||_2{}^2 := w_1^2 + w_2^2 + \cdots + w_d^2$ penalizes the parameter vector being too large, and it is known to work well when we predict with data outside the given data set.

When the dimensionality of the feature vector is extremely large compared with the number of data, the problem called over-fitting arises, which is the phenomenon that the predictive performance for the new data other than the data used for fitting the model severely degrades.

In this article, we predicted the existence of citation that already existed because the aim of this article is link (citation) modelling, which discovers the factors affecting the existence of citations. However, of course, in the following experiments, we hid the existing links and predicted them.

Feature Selection

We designed features in the citation network as shown in Table 1. We categorized 11 features into three types: four topological features, one semantic feature, and six attribute features. In this article, all of the features are the scores of pairs of two papers. The detailed definitions of features of each citation pair (from a paper *from* to another paper *to*) are described below.

Topological Features

(1) The number of common neighbours. In many social networks in the real world with *small-world* phenomenon, nodes are highly clustered locally (Watts & Strogatz, 1998), which means two papers with more common neighbours tend to be connected. The number of common neighbours is defined as the number of nodes that is connected to both *from* and *to*. When we count the number of common neighbours, we regard all citations in as undirected link, such as from Figure 1 (A) to (B). In the example of Figure 1, the common neighbours of a pair *(F, T)* are node X, Y, and Z. We hypothesize, as we observe in the social networks in our life, that the more common the neighbours are, the more likely a citation exists.

(2) Link-based Jaccard coefficient. The link-based Jaccard coefficient is the size of the intersection divided by the size of the union of the neighbours of *from* and *to* and defined as:

$$J_{link}(from, to) = \frac{|L_{from} \cap L_{10}|}{|L_{from} \cup L_{10}|}$$

where L_x represents the set of links (citations) of document x. The link-based Jaccard coefficient represents the relative value of the number of common neighbours. In the example of Figure 1, the link-based Jaccard coefficient of a link (F, T) is $3 / 7 = 0.43$. The link-based Jaccard coefficient is similar to but different from the clustering coefficient by Watts and Strogatz (1998). The clustering coefficient is a value of a node, while the link-based Jaccard coefficient represents the status of a link (an edge).

(3) Difference in betweenness centrality. A previous study revealed that betweenness centrality correlated with the citations expected in the distant future (Shibata, Kajikawa, & Matsushima, 2007). Betweenness centrality represents the extent to which a node lies on the paths between other nodes and can also be interpreted as measuring the influence a node has over the spread of information through the network. A paper with a large betweenness centrality bridges unconnected papers and is therefore anticipated as a previously

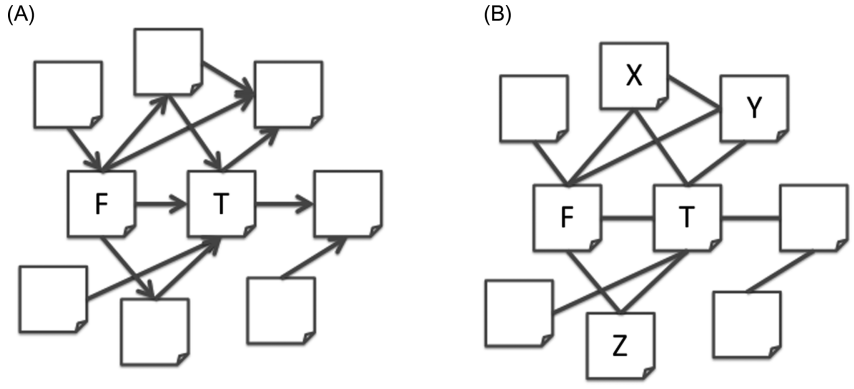

FIG. 1. An example of a citation network.

DOI: 10.1002/asi

unexplored seed of innovation. The betweenness centrality of node i, *Bc(i)* is given by following step:

1. To pick up a pair of nodes s and t other than i.
2. To count the number of shortest paths (σ_{st}) between s and t.
3. To count the number of shortest paths ($\sigma_{st}(i)$) between s and t through i.
4. To calculate the ratio by $\sigma_{st}(i)/\sigma_{st}$.
5. To repeat Steps 1 to 4 for all pairs of s and t and sum up $\sigma_{st}(i)/\sigma_{st}$.

Formally, the betweenness centrality of node *i, Bc(i)* is defined as:

$$B_c[i] = \sum_{s \neq i \neq t \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

where σ_{st} is the total number of shortest paths from node s to node t, and $\sigma_{st}(i)$ is the number of shortest paths from s to t traversing i (Freeman, 1977). The definition of the difference in betweenness centrality is defined as:

$$dif_{bc} = Bc(to) - Bc(from)$$

where *Bc(i)* represents the betweenness centrality of node i in the citation network. In this article, we calculate the difference in betweenness centralities of *to* and *from* to represent a relational value. For instance, assume that a certain paper is "center" with large betweenness centrality and another paper is in the "leaf" with very small betweenness centrality. Our hypothesis is the large gap of betweenness centralities will affect the existence of citations.

(4) Difference in the number of in-links. As Barabási et al. (2002) and Newman (2001) have further proposed, on the basis of empirical evidence, that the probability of a certain paper to be cited is correlated with the number of citations. The definition of the difference in the number of in-links is defined as:

$$dif_{in-link} = \#inlink(to) - \#inlink(from)$$

Similar to the previous one, we calculate the difference in the number of in-links of *to* and *from* to represent a relational value. This value represents the attracting force between *from* and *to*, while the absolute value of citation toward *to*, (11) *number of times "to" cited* represents the gravity.

(5) Is same cluster. After extracting the largest connected component, the citation network is divided into clusters using the topological clustering method (Newman, 2004), which is not fuzzy. Newman's algorithm discovers tightly knit clusters with a high density of within-cluster edges, which enables the creation of a nonweighted graph that comprises many nodes. This value is whether the two nodes, *from* and *to*, are classified into the same cluster. The value is 1 if the same, and 0 otherwise. We hypothesize that two papers are more likely to be connected if they are in the same clusters.

Semantic Features

(6) Cosine similarity of term frequency–inverse document frequency (tf–idf) vectors. Cosine similarity of *tf–idf* vectors can be calculated as the inner product of the *tf–idf* vectors of two documents and is defined as:

$$\cos ine(from, to) = \vec{w}_{from} \cdot \vec{w}_{to} = \sum_i W_{from}{}^{(i)} W_{to}{}^{(i)}$$

where $w_d{}^{(i)} = tf_{i,d} \times \log\left(\frac{N}{df_i}\right)$, $tf_{i,d}$ represents #occurrences of i-th term in document d, df_i represents #document containing i-th term, and N represents total number of documents. In the calculation of similarity measures, nouns and nominal phrases are extracted from each abstract by linguistic filtering to represent a paper as a term vector (Frantzi et al., 2000). We hypothesize that the more similar the topics of two papers are, the more likely that a citation exists.

Attribute Features

(7) Difference in publication year. Papers tend to cite recent papers to show that they are making an original and timely contribution. Therefore, a difference in publication year has a negative effect on citedness. The difference in the publication year is defined as:

$$dif_{year} = published(to) - published(from)$$

(8) The number of common authors. The number of common authors represents the number of common authors between *from* and *to*. The number of common authors is expected to have a similar effect to that of self-citation.

(9) Is self citation. This value represents whether the two papers are written by at least one author. MacRoberts and MacRoberts (1989) pointed out that self-citations are often observed because researchers are more familiar with their own research than the research of others. The value is 1 if self-cited, and 0 otherwise.

(10) Is published in same journal. This value is whether the two papers are published in the same journal. The value is 1 if the same, and 0 otherwise. As researchers tend to work in specific scientific communities, two papers published in the same journal are likely to be connected.

(11) Number of times "to" cited. The number of times "*to*" is cited represents the number of in-links of "*to*." Preferential attachment has received considerable attention as a model of network growth (Newman, 2003). In the preferential attachment model, the probability of a new edge involving node x is proportional to the number of links of x. With this model, the node with more citations will receive more citations, where rich get richer.

DOI: 10.1002/asi

TABLE 2. Datasets of citation networks.

Dataset	Query	Published through	No. of papers	No. of citations
A Innovation	innovation*	2009	20,564	106,619
B Nano Bio	nano* and bio*	2009	33,830	175,875
C Organic LED	((organic* or polymer*) and (electroluminescen* or electro-luminescen* or electro luminescen* or light emitting or LED*)) or OLED*	2009	19,486	196,123
D Solar Cells	solar cell*	2008	18,587	111,051
E Secondary Batteries (*)	((secondary or storage or rechargeable or reserve) and cell*) or batter*	2008	20,430	145,008

Data and Experiment

In this article, five large-scale citation datasets, Innovation, Nano Bio, Organic LED, Solar Cells, and Secondary Batteries, are collected as shown in Table 2. We searched databases of academic papers and patents using the same query for each domain. The databases of academic papers used are the Science Citation Index Expanded (SCI-EXPANDED), the Social Sciences Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI) compiled by the Institute for Scientific Information (ISI). After collecting data, we extracted the papers and citations in the largest-graph component to eliminate those not linked to any other papers. The detailed citation analysis of Innovation, Nano Bio, Organic LED, Solar Cells, and Secondary Batteries are found respectively in Hashimoto, Kajikawa, Sakata, Takeda, and Matsushima (in press), Takeda, Mae, Kajikawa, and Matsushima (2009), Kajikawa and Takeda (2009), Kajikawa, Yoshikawa, Takeda, and Matsushima (2008), and Shibata, Kajikawa, and Sakata (2010). In the case of secondary batteries, we filter only the papers related to batteries from papers in other research areas in biomedicine following the procedure applied in previous research (Shibat, Kajikawa, and Sakata, in press). According to our analysis, a huge number of bio-related researches were included in the original corpus especially after 1990. Based on our experiences, the single query search will extract more documents we should collect. Our strategy to collect documents was to avoid missing documents that we should have included in a certain research domain, but then we filtered by using only the largest connected component.

As training data, we extracted tens of thousands of citation pairs from the citation network and used these pairs as positive instances. As negative instances, we randomly selected the same number of pairs of papers that do not have citation relations. We created five sets of the training data with five different sets of negative instances. For instance, in the case of Nano Bio in Table 2, there were 106,619 citations. Our procedures were as follows:

1. These existing citations are divided into five groups (positive instances, namely, P[1] to P[5]).
2. We randomly created the same number of pair where citations did not exist (negative instances, namely, N[1] to N[5]).
3. In the first experiment, P[2] to P [5] and N[2] to N[5] were used as the training data and P[1] and N[1] were used as the test data.
4. We repeated step 3 five times in total with different choice of answer set.

TABLE 3. Prediction results.

Dataset	Precision	Recall	F1
A Innovation	0.75	0.91	0.82
B Nano Bio	0.83	0.76	0.79
C Organic LED	0.79	0.71	0.74
D Solar Cells	0.76	0.72	0.74
E Secondary Batteries	0.80	0.77	0.77

As a learner, we employed L2-regularized and L2-loss SVM. As we saw in the formulation of the SVM, the regularization term plays a role in avoiding the over-fitting problem. As the regularization term, we employ L2- regularization and L2-loss, $||w||_2^2$. The optimal parameter for the regularization was chosen with a greedy search. We measured the performance of L2-regularized and L2-loss SVM by using five-fold cross validation.

Results

The performance of L2-regularized and L2-loss SVM is shown with precision, recall and F1 values in Table 3. The F1 values were between 0.74 and 0.82. Based on the results, we obtained the learning model on our training data by using L2- regularized and L2-loss SVM. The model includes 11 weight features. Table 4 shows the weights of all features obtained from the model with the best F1 value in each research domain. We judged the contribution based on the weight in Table 4. Positive contributions mean the weight of equal or more than 0.5 and negative ones are equal or less than −0.5. (1) *The number of common neighbours*, (2) *link-based Jaccard coefficient*, (3) *difference in betweenness centrality*, (6) *cosine similarity of tf–idf vectors*, (9) *is self citation*, and (10) *is published in the same journal* contributed to the prediction of citations. On the other hand, (4) *difference in the number of in-links*, (5) *is same cluster*, (7) *difference in publication year*, (8) *the number of common authors*, (11) *number of times "to" cited* did not contribute to the predictions.

Discussion

Our learner performed very well with the F1 values of between 0.74 to 0.82, which means our model, can predict

DOI: 10.1002/asi

TABLE 4. Weights of features.

Features	A. Innovation	B. Nano Bio	C. Organic LED	D. Solar Cells	E. Secondary Batteries
1. No. common neighbors	0.566	0.889	0.520	0.683	0.987
2. Link-based Jaccard coefficient	1.354	2.198	−6.150	−0.703	−4.742
3. Difference in betweenness centrality	−1.446	−6.107	−2.175	−5.468	−10.049
4. Difference in the number of in-links	0.052	0.033	0.034	0.045	0.047
5. Is same cluster	0.018	0.086	−0.308	−0.160	−0.062
6. Cosine similarity of tf-idf vectors	−19.897	−17.817	−15.527	1.624	1.519
7. Difference in publication year	0.018	0.046	0.032	0.009	0.008
8. The number of common authors	−0.112	0.476	0.403	0.152	0.036
9. Is self-citation	1.975	0.756	0.605	0.865	0.918
10. Is published in same journal	0.726	0.614	0.198	0.027	−0.108
11. Number of times "to" cited	−0.018	−0.019	−0.015	−0.031	−0.033

the existence of a citation with a probability from 74% to 82%, given a pair of papers and the entire citation network. Especially three features, (2) *link-based Jaccard coefficient*, (3) *difference in betweenness centrality*, and (6) *cosine similarity of tf–idf vectors*, largely affected the predictions of citations.

Link-based Jaccard coefficient contributed positively in the cases of (A) Innovations (weight: 1.354) and (B) Nano Bio (2.198) but negatively in the cases of (C) Organic LED (−6.150), (D) Solar Cells (−0.703) and (E) Secondary Batteries (−4.742). These results indicate that the former research areas, such as (A) Innovations and (B) Nano Bio, comprise multiple research fields and most citations are in each research field so that papers cite locally, i.e., in each research field. On the other hand, (C) Organic LED, (D) Solar Cells, and (E) Secondary Batteries are contained in a research field with a single issue so that citations are global and each pair of papers tends to have a relatively small number of common neighbors in the citation networks.

Difference in betweenness centrality had negative impacts in all cases, which indicates that the larger *Bc(from)* is than *Bc(to)*, the more likely the citation is to exist. This result is understandable because it is rare that core nodes and peripheral nodes, which have different values of betweenness centrality, are linked in the citation networks.

Cosine similarity of tf–idf vectors affected (A) Innovations (−19.897), (B) Nano Bio (−17.817), and (C) Organic LED (−15.527) negatively, while (D) Solar Cells (1.624) and (E) Secondary Batteries (1.519) did so positively. This can be interpreted as follows. The first three research areas contain various research seeds. For instance, there have been many candidate materials, such as gallium nitride (GaN), zinc selenide (ZnSe), and zinc oxide (ZnO) for organic LED. In these cases, overlaps of terms with large cosine similarity do not generate citations. On the contrary, the last two research fronts have relatively fewer technological seeds so that a certain paper tends to cite another one if they have common terms in their descriptions.

In addition to the three features described above, another three features affected the existence of citations. *The number of common neighbours* positively affected all cases, because the more common neighbours two papers have, the more related they are. That the *self-citation* result had a positive effect is reasonable because authors tend to cite their own papers. The feature of *is published in the same journal* affected only (A) Innovations (0.726) and (B) Nano Bio (0.614) positively. Similar to the result of *link-based Jaccard coefficient*, papers tend to cite in each research field in the case of research fields with multiple issues.

In summary, different models are required for different types of research areas—research fields with a single issue or research fields with multiple issues. In the case of research fields with multiple issues, there are barriers among research fields because our results indicate that papers tend to be cited in each research field locally. Therefore, one must consider the typology of targeted research areas when building models for link prediction in citation networks. It is difficult to build a universal learner for link prediction and we need to build learners based on the characteristics of each research domain. We can see two types of models from our experiments. The first one is the research field with multiple issues such as (A) Innovations and (B) Nano Bio. The second one is a simple research field type with commonly understood targets of research and development such as (C) Organic LED, (D) Solar Cells and (E) Secondary Batteries.

There are two remaining limitations to be tackled in future studies. The first is to consider the dynamics. In this article, we have not considered the dynamical evolution of citation networks because our primary purpose of this research is to build models for link prediction in citation networks. However, if author(s) or reviewers of a certain research article would like to extract related papers to be cited before the paper is published, then we need to predict the existence of links given a paper published in the year $t + 1$ and a citation network up through the year t. In this case, some of the topological features would not be available, as the paper has not been published yet.

Second, researchers or reviewers should use the outcomes of link predictions as candidates of lists of references. Although the outputs suggest lists of publications possibly to

DOI: 10.1002/asi

be cited, the decision whether to cite these suggested papers should be made by human beings. Although our method can suggest lists of publications that can be missed being cited, research outcomes should be created based on the intelligence of researchers. When such an expert judgment is iteratively included in our model, the accuracy will be improved.

Finally, as there is a strong collineality between (2) *Link-based Jaccard coefficient*, (5) *is same cluster*, and (6) *cosine similarity of tf–idf vectors*, the weight of (2) *Link-based Jaccard coefficient* will increase if we remove (6) *cosine similarity of tf–idf vectors*. The reason (2) *Link-based Jaccard coefficient* had a greater affect than (5) *is same cluster* was that citations are locally dense not globally. This feature selection is our future study because we need to be careful to add/remove features to compare.

Conclusion

We built models to predict the existence of citations among papers by formulating link predictions for five large-scale datasets of citation networks. A supervised machine-learning model was applied with 11 features. As a result, our learner performed very well with F1 values of between 0.74 and 0.82. Three features in particular, *link-based Jaccard coefficient*, *difference in betweenness centrality*, and *cosine similarity of tf–idf vectors*, largely affected the predictions of citations. The results also indicated that different models are required for different types of research areas—research fields with a single issue or research fields with multiple issues. In the case of research fields with multiple issues, there are barriers because papers tend to be cited in each research field locally. Therefore, one must consider the typology of targeted research areas when building models for link predictions in citation networks.

Acknowledgment

One of the authors (YK) was partially supported by the New Energy and Industrial Technology Development Organization (NEDO), Grant for Industrial Technology Research (09D47001a). YK was also supported by the Ministry of Education, Science, Sports and Culture (MEXT), Grant-in-Aid for Young Scientists (B) (21700266).

References

Barabási, A., Jeong, H., Ne 'da, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaboration. Physica A, 311, 3–4.

Börner, K., Chen, C.M., & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology, 37, 179–255.

Boyack, K.W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. Scientometrics, 64(3), 351–374.

Boyack, K.W., Wylie, B.N., & Davidson, G.S. (2002). Domain visualization using VxInsight for science and technology management. Journal of the American Society for Information Science and Technology, 53, 764–774.

Caragea, D., Bahirwani, V., Aljandal, W., & Hsu, W.H. (2009). Ontology-based link prediction in the LiveJournal Social Network. Proceedings of the Eighth Symposium on Abstraction, Reformulation, and Approximation (SARA2009).

Chen, C. (1999). Visualizing semantic spaces and author co-citation networks in digital libraries. Information Processing and Management, 35(2), 401–420.

Chen, C., Cribbin, T., Macredie, R., & Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: two case studies. Journal of the American Society for Information Science and Technology, 53, 678–689.

Clauset. A, Moore, C., & Newman, M.E.J. (2008). Hierarchical structure and the prediction of missing links in networks. Nature, 453, 98–101.

Davidson, G.S., Hendrickson, B., Johnson, D.K., Meyers, C.E., & Wylie, B.N. (1998). Knowledge mining with VxInsight: Discovery through interaction. Journal of Intelligent Information Systems, 11, 259–285.

de Solla Price, D.J. (1965). Networks of scientific papers. Science, 149, 510–515.

Freeman, L.C. (1977). A set of measures of centrality based on betweenness, Sociometry, 40, 35–41.

Hashimoto, M., Kajikawa, Y., Sakata, I., Takeda, Y., & Matsushima K. (in press). Academic landscape of innovation research and National Innovation System policy reformation in Japan and the United States. International Journal of Innovation and Technology Management.

Hasan, M.A., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. SIAM 2006 Workshop on Link Analysis, Counterterrorism and Security, Bethesda, MD.

Hwang, S.Y., Wylie, B.N., Wei, C.P., & Liao, Y.F. (2010). Coauthorship networks and academic literature recommendation. Electronic Commerce Research and Applications, 9, 323–334.

Kajikawa, Y., Abe, K., & Noda, S. (2006). Filling the gap between researchers studying different materials and different methods: A proposal for structured keywords. Journal of Information Science, 32, 511–524.

Kajikawa, Y., & Takeda, Y. (2009). Citation network analysis of organic LEDs, Technological Forecasting and Social Change, 76, 1115–1123.

Kajikawa, Y., Yoshikawa, J., Takeda, Y., & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy, Technological Forecasting and Social Change, 75, 771–782.

Klavans, R., & Boyack, K.W. (2009). Toward a consensus map of science. Journal of the American Society for Information Science and Technology, 60(3), 455–476.

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. In Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM) (pp. 556–559). New York: ACM Press.

Lü, L., Jin, C.H., & Zhou, T. (2009). Similarity index based on local paths for link prediction of complex networks. Physical Review E, 80, 046122.

MacRoberts, M.H., & MacRoberts, B.F. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40(5), 342–349.

Murata, T., & Moriyasu, S. (2008). Link Prediction based on Structural Properties of Online Social Networks. New Generation Computing, 26(3), 245–257.

Newman, M.E.J. (2001). Clustering and preferential attachment in growing networks. Physical Review E, 64, 025102.

Newman, M.E.J. (2003). The structure and function of complex networks. SIAM Review, 45, 167–256.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69, 066133.

Popescul, A., & Ungar, L.H. (2003). Statistical relational learning for link prediction. Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003.

Seglen, P.O. (1994). Casual Relationship between article citedness and journal impact. Journal of the American Society for Information Science and Technology, 45(1), 1–11.

Shibata, N., Kajikawa, K., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core papers. Journal of the American Society for Information Science and Technology, 58 (6), 872–882.

Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Assessing the gap between science and technology: Case study of secondary batteries. Nihon Chizai Gakkaishi, 6, 5–12.

Shibata, N., Kajikawa, Y., & Sakata, I. (in press). Detecting potential technological fronts by comparing scientific papers and patents. Foresight.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, 28(11), 758–775.

Small, H. (1999). Visualizing science by citation mapping. Journal of the American Society for Information Science and Technology, 50(9), 799–813.

Small, H. (2003). Paradigms, citations, and maps of science: A personal history. Journal of the American Society for Information Science, 54(5), 394–399.

Takeda, Y., Mae, S., Kajikawa, Y., & Matsushima, K. (2009). Nanobiotechnology as an emerging research domain from nanotechnology: A bibliometric approach, Scientometrics, 80, 23–38.

Vapnik, V.N. (2000). The nature of statistical learning theory. New York: Springer.

Vinkler, P., & Davidson, G.S. (2002). Dynamic changes in the chance for citedness. Scientometrics, 54(3), 421–434.

Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. Nature, 393, 440–442.

DOI: 10.1002/asi