As academic disciplines are segmented and specialized, it becomes more difficult to capture relevant research areas precisely by common retrieval strategies using either keywords or journal categories. This paper proposes a method of measuring the relatedness among sets of academic papers in order to detect unrelated communities which are not related to target topic. A citation network, extracted by given keywords, is divided into communities based on the density of links. We measured and compared four measures of relatedness between two communities in a citation network for three large-scale citation datasets. We used both link and semantic similarities. The topological distance from the center in a citation network is a more efficient measure for removing the unrelated communities than the other three measures: the ratio of the number of intercluster links over the all links, the ratio of the number of common terms over all terms, cosine similarity of tf-idf vectors.

Introduction

Since the number of academic papers exponentially increases (Price, 1965), each academic area becomes specialized and segmented. Davidson, Hendrickson, Johnson, Meyers, and Wylie (1998, p. 259) reflect on 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.” 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 a narrower domain than ever. Therefore, segmentation occurs simultaneously with specialization, which brings a severe problem and also opportunity to find crucial knowledge by integrating different domains (Kajikawa, Abe, & Noda, 2006).

Because this trend of specialization and segmentation and also information flood is the nature of science, there is a strong need for computational tools of science mapping and emerging topic detection. Previous studies have established smart algorithms for creating academic landscapes and for detecting emerging topics for certain research fronts. Recently, methods of science mapping by citation analysis and text mining have been proposed (Small, 2003; Boyack, Klavans, & Börner, 2005; Klavans & Boyack, 2009). Researchers have also focused on clustering and visualization (Chen, 1999; Small, 1999; Börner, Chen, & Boyack, 2003). Leydesdorff and colleagues made a large-scale investigation of a set of academic papers (Leydesdorff, 2004; Leydesdorff & Rafols, 2009). Not only creating static academic landscapes, topological and semantic analysis of a citation network also helps us to focus on significant movements in research fronts and emerging research fields in a broad context (Boyack, Wylie, & Davidson, 2002; Chen, Cribbin, Macredie, & Morar, 2002; Shibata, Kajikawa, Takeda, & Matsushima, 2008).

Moreover, in such circumstances where a single segmented discipline cannot solve complex social issues, many opportunities exist among distant disciplines, and therefore interdisciplinary research is expected to bridge different conventional disciplines and to open up new avenues of research. In an earlier paper, Kajikawa et al. (2006) illustrated that even within a discipline research is segmented into different research domains and we can discover new knowledge by combining distant domains. The significance of combining outcomes from different disciplines has been pointed out (Piccoli & Ives, 2005). Haynes, Hayward, and Lomas (1995) explained how the knowledge of basic science was applied to real clinical problems. Previous studies have focused on...
the information process in interdisciplinary research, investigating how collaborations among different disciplines occur (Qin, Lancaster, & Allen, 1997), extracting factors for determining the transformation of knowledge (Wilson, 1997) and analyzing the structural and strategic elements of information exchange between intellectual domains (Palmer, 1999). Klavans and Boyack (2006) proposed a measure of relatedness between bibliometric units, such as journals, documents, authors, or words. These research efforts have the potential to extract the opportunity by measuring the relatedness of different domains, which is currently hidden between domains.

When the amount of information is not so large, researchers can find such an opportunity among domains by deliberately investigating existing papers; however, as the amount of publications rapidly increases with unprecedented pace and the relationships among disciplines become more complex than experts can comprehend, the problem of how to extract the knowledge domain to be analyzed without noise becomes more serious. The number of published papers and its trend are generally used for evaluating the importance of a research area; however, how can we retrieve and determine a set of papers representing that area? There are two major ways of retrieving papers from a given research area. The method is to collect all papers published in journals of certain journal categories (Leydesdorff, 2004; Leydesdorff & Rafols, 2009). Journal categories can extract papers precisely if the target matches conventional journal categories exactly but does not work when the granularity of the target is larger or smaller than the journal categories and is an interdisciplinary one.

The other method is to use keywords as a query and search papers matching the query (Kajikawa, Ohno, Takeda, Matsushima, & Komiyama, 2007). Queries are selected according to the following two steps: (1) the representative keyword is selected but (2) if the definition of its domain is unclear, more keywords are added. The second step is called query expansion (Kostoff, Eberhart, & Toothman, 1997). In order to retain wide coverage of a research area and to avoid omission of core papers, we need to use more terms as a query, but then the results contain more papers unrelated to the topics. Once the entire papers are extracted by journal categories or keywords, papers are divided into several communities based on the structure of its citation network. Then we compare and discuss four measures of the distances between two communities. Two of four are link-based measures and the others are text-based.

**Methodology and Datasets**

**Methodology**

The entire procedure is described in Figure 1. Step (1) involves collecting academic papers from a database. We search databases of academic papers by queries. The databases of the academic papers used are the Science Citation Index Expanded (SCI-EXPANDED), the Social Sciences Citation Index (SSCI), and the Arts and Humanities Citation Index (A&HCI) compiled by the Institute for Scientific Information (ISI), because SCI-EXPANDED, SSCI, and A&HCI are three of the best sources for citation data. This database enabled us to obtain both the attribute data of each document such as the year published, title, author(s), abstract, and so forth, and relational data, i.e., citation data.

Subsequently, as Step (2), we constructed citation networks by regarding papers as nodes and citations as links. According to a previous study, intercitation, which is also sometimes known as direct citation, is the best way of detecting emerging trends and can work well even when the size of the corpus is small (Shibata, Kajikawa, Takeda, & Matsushima, 2009). In network analysis, only the data of the largest-graph component are analyzed, because this paper focuses on the relationship among papers, and we should therefore eliminate those not linked with any other papers in Step (3).

After extracting the largest connected component, in Step (4) the network is divided into communities using the topological clustering method (Newman, 2004). Newman’s algorithm discovers tightly knit communities with a high density of within-community edges, which enables the creation of a nonweighted graph consisting of many nodes. The algorithm proposed is based on the idea of modularity. Modularity $Q$ is defined as follows:

$$Q = \sum_s (e_{ss} - a_s^2) = Tr(e) - \|e\|^2$$

where $e_{ss}$ is the fraction of the edges in the network that connect nodes in cluster $s$ to those in cluster $t$, and $a_s = \sum_t e_{st}$. The first part of the equation, $Tr(e)$, represents the sum of
the density of edges within each cluster. A high value of this parameter means that nodes are densely connected within each cluster. However, the maximum value of this \((Tr(e) = 1)\) is given if all nodes are regarded as one cluster. The second part of the equation, \(\|e\|^2\), represents the sum of the density of edges within each cluster when all edges are placed randomly. Then the algorithm to optimize \(Q\) to gain all possible divisions to find the best structure of clusters is as follows. The algorithm does not allow overlaps of clusters. Starting with a state in which each node is the only member of one of the \(n\) clusters, we repeatedly join clusters together in pairs, choosing at each step the join that results in the greatest increase in \(Q\). The change in \(Q\) upon joining two clusters is given by:

\[
\Delta Q = e_{st} + e_{ts} - 2a_{st}a_{ts} = 2(e_{st} - a_{st}a_{ts})
\]

The iterations are stopped when \(\Delta Q\) becomes negative, because the purpose here is not to gain a whole dendrogram but extract more relevant structures in regard to citation networks. Takeda and Kajikawa (2010) showed that this algorithm can effectively extract research communities when the size of the community is larger than several tens of nodes, while it will become noisy when the size is less than a few tens of nodes.

After clustering, as Step (5), the distance between two communities is measured for each pair of communities. The aim is to detect communities far from the targeted research topic, given the division of communities in a citation network. The basic strategy is as follows; (1) for each community \(c\), find the nearest community \(t\) by calculating the relatedness between \(s\) and \(t\), and (2) if the relatedness between \(s\) and \(t\) is much smaller than that of the others, \(s\) can be regarded as an unrelated community. There are two main ways to measure the relatedness between two sets of documents, topological relatedness (network coherence) and semantic relatedness (textual coherence). In this paper we calculate four relatedness measures, which are defined below. Finally the predictions based on these four measures are compared to the judgments by experts. In the following, we show the definition of each measure.

(A) Link-based Jaccard coefficient: \(J_{link}\). It is expected that the denser the links between two communities are, the closer the two communities are. The density between two communities can be measured by the size of the intersection divided by the size of the union. The link-based Jaccard coefficient is the number of the intercommunity edges divided by the sum of the number of edges of both communities. The link-based Jaccard coefficient between community \(c_1\) and \(c_2\) is defined as:

\[
J_{link}(c_1, c_2) = \frac{2 \times I(c_1, c_2)}{E_{c_1} + E_{c_2}}
\]

where \(E_r\) represents the total number of edges of the nodes in community \(c\) and \(I(y, z)\) is the number of intercommunity edges between \(y\) and \(z\).

If a certain community has fewer edges compared to those of all the other ones, the community is far from the targeted research area. Therefore, we define the maximum link-based Jaccard coefficient of \(c_1\) and the corresponding community as:

\[
J_{link}(c_1) = \max_{c_2} J_{link}(c_1, c_2),
\]

\[
\arg \max_{c_2} J_{link}(c_1) = \arg \max_{c_2} J_{link}(c_1, c_2)
\]

If \(J_{link}\) of a certain community is significantly smaller than those of other communities, this community is assumed to be unrelated to the targeted research topic.

(B) The average shortest path from the center of the graph: \(L\). The more paths the nodes in a certain community take from the center, the farther the community is from the targeted research area. The average shortest paths from the center of community \(c\) is measured by the average shortest paths from the closeness center in the given graph among the nodes in community \(c\), and defined as:

\[
L(c) = \frac{1}{|z|} \sum_{i \in c} d(i, z)
\]

where \(d(i, j)\) is the shortest paths between \(i\) and \(j\), and \(z\) represents the center node in the given graph in terms of the average paths to all other nodes (Sabidussi, 1966; Watts & Strogatz, 1998). If \(L\) of a certain community is significantly larger than those of other communities, this community is assumed to be unrelated to the targeted research topic.

(C) Jaccard coefficient of term vectors: \(J\). The above two measures, i.e., \(J_{link}\) and \(L\), are based on the network coherence. The following represent textual coherence. Our assumption is that the more common terms two communities share, the closer the two communities are. Terms are extracted by linguistic filtering, using abstracts of papers (Mima, Frantz, & Ananiadou, 1998; Frantz, Ananiadou, & Mima, 2000). Linguistic filtering extracts candidate noun phrases, such as:

1. Noun + Noun,
2. (Adj | Noun) + Noun,
3. ((Adj | Noun)* (NounPrep)?)(Adj | Noun)* Noun.

Then the relatedness between two communities can be measured by the size of textual intersection divided by the size of text in the union. The Jaccard coefficient between community \(c_1\) and \(c_2\) is defined as:

\[
J(c_1, c_2) = \frac{|T_{c_1} \cap T_{c_2}|}{|T_{c_1} \cup T_{c_2}|}
\]

where \(T_r\) represents the set of nouns or nominal phrases in the documents belonging to community \(c\) (Tan, Steinbach, & Kumar, 2005). If a given community pair shares few terms with all the other ones, the community is far from the targeted research area. Therefore, we define the maximum Jaccard coefficient of \(c_1\) and the corresponding community as:

\[
J_{c_1} = \max_{c_2} J(c_1, c_2),
\]

\[
\arg \max_{c_2} J(c_1) = \arg \max_{c_2} J(c_1, c_2)
\]
If $J$ of a certain community is significantly smaller than the ones of other communities, this community is assumed to be unrelated to the targeted research topic.

(D) Cosine similarity of tf-idf vectors: $C$. This measure assumes that the more common terms two communities share, the closer the two communities are. The relatedness between two communities can be measured by the cosine similarity (i.e., inner dot product) of $tf$-$idf$ vectors. Cosine similarity of $tf$-$idf$ vectors between community $c_1$ and $c_2$ is defined as:

$$C(c_1, c_2) = \cosine(c_1, c_2) = \frac{w_{c_1}^T \cdot w_{c_2}^T}{\|w_{c_1}\| \cdot \|w_{c_2}\|}$$

where $w_{c_i} = tf_{i,c} \times \log \frac{N}{df_i}$: the number of occurrences of $i_{th}$ term in community $c$, $df_i$ is the number of documents containing $i_{th}$ term and $N$ is the total number of documents (Tan et al., 2005). Note that all $w_c$ values are normalized as $\|w_c\| = 1$. Usually $tf$-$idf$ is a value of a term in a certain document, but, in this paper, we extend it into a value of a term in a certain community (a set of documents).

If a given community pair shares few terms with all other ones, the community is far from the targeted research area. Therefore, we define the maximum cosine similarity of $tf$-$idf$ vectors of $c_1$ and the corresponding community as:

$$C(c_1) = \max_{c_2} C(c_1, c_2),$$

$$\arg \max[C(c_1)] = \arg \max C(c_1, c_2)$$

If $C$ of a certain community is significantly smaller than the ones of other communities, this community is assumed to be unrelated to the targeted research topic.

Datasets

We evaluate four measures with three large-scale citation data as shown in Table 1. The targets of research domains are nanobiotechnology (NBT), organic light emitting devices (OLED), and innovations (INV). While NBT and OLED are purely technological subjects, the papers in INV combine many disciplines.

NBT is an emerging research area within nanotechnology, which is defined as a field that applies nanoscale principles and techniques to understanding and transforming biosystems (living or nonliving) and one that uses biological principles and materials to create new devices and systems integrated from the nanoscale (Roco, 2003). However, it is not an easy task to define NBT and collect relevant papers. In a previous bibliometric study, the set of query nano* and bio* was used (Takeda & Kajikawa, 2010), but it contains some papers that can be considered noise in the retrieval process. We searched papers using nano* and bio* as shown in Table 1.

OLED is also an emerging research area, and covers new applications in the semiconductor industry where the basic components of the industry are thin metal and semiconductor films fabricated on silicon substrates. The role of polymers in the electronics industry is conventionally associated with insulating properties, but some organic materials are expected to be realized as key devices in the electronics industry. We use the following query ((organic* OR polymer*) AND (electroluminescence* OR electro-luminescence* OR light emitting OR LED*)) OR OLED*, used in the previous research (Kajikawa & Takeda, 2009). This query can retrieve a high coverage of relevant papers, but includes noisy terms because of its multiple terms.

INV is the process of bringing new products and services to market, and is one of the most important topics not only in management research, but also in economics, policy, engineering, and so forth. Innovation is a broad topic, and a variety of disciplines such as engineering, business, management, economics, and sociology address various aspects of innovation, including technological innovation, business innovation, social innovation, and circumstances for incubating innovations (Gopalakrishnan & Damanpour, 1997). Although we use the simple query, innovation*, there is a possibility that the retrieved corpus includes papers that can be considered noise in the retrieval process because innovation has another connotation in the study of time series analysis. In time series analysis, a set of past data is used to predict the future. When we can develop the relevant model which described well past data, we can usually predict the future with that model. This is based on the assumption that the mechanism governing the data does not change in the future. However, we sometimes observe data at a certain timepoint which does not obey the model. In the studies of time series analysis, that point is called innovation.
In this paper we focus on the top 15 communities of the number of papers. In all research domains, the top 15 communities contain more than 95% of papers in the largest connected components. Based on the four measures, unrelated communities were predicted and then the results were compared to the judgments by experts of each research domain.

Results

The detailed results of the top 15 communities in each research area are described in Table 2. In Table 2, the average publication year can be calculated as the sum of publication year divided by the number of papers in each community. The topics in Table 2 are extracted by experts in each research domain from the top 10 papers based on the number of citations in each community. The main objective of NBT is to synthesize and analyze nanomaterials with organic materials whose size is on the order of nanometer and can be applied in bioimaging and biosensors. For example, the biggest cluster in NBT (NBT #1) analyzed nanostructures of nanobio materials. The second largest cluster in NBT (NBT #2) studies an application of NBT in bioimaging. However, there are unrelated communities like nanoplankton (NBT #6) and nanobacteria (NBT #12). The details of the unrelated communities are discussed later. In OLED, researchers study materials, device structure, performance, and processing. However, some communities study the polymerization reactions under the luminescence of LEDs and other topics relating inorganic LEDs but not OLEDs. In INV, we can also see unrelated issues like genome and evolution (INV #6) and time-series analysis (INV #7) where the term innovation is used in other connotations, as described earlier.

Relationships between the rank of community size and four relatedness measures are plotted in Figure 2. White circles represent the communities judged unrelated by experts, while black triangles are the communities that are relevant with the same notations. For methods (A), (B), (C), and (D) represent the link-based Jaccard coefficient, \( J_{\text{link}} \), the average shortest path from the center of the graph, \( L \), Jaccard coefficient of term vectors, \( J \), and cosine similarity of \( tf-idf \) vectors, \( C \), respectively. The value of these measures is plotted according to the vertical axis. The horizontal axis represents the rank in the order of community’s size, i.e., the number of papers in each community. In the following figures and tables (1), (2), and (3) represent the analysis of NBT, OLED, and INV, respectively.

Similar trends can be observed in all three cases for each measure. Figure 2A, C, D indicate that \( J_{\text{link}} \), \( J \), and \( C \) could not distinguish unrelated communities from others by a horizontal line. Especially for semantic measures, \( J \) and \( C \), all plots seem to be represented by one line or curve, which means that we cannot distinguish related and unrelated communities by these measures. However, \( L \) values seem to have the potential to distinguish the unrelated from related ones with few errors.

As shown in Figure 2A-1, A-2, A-3, the unrelated communities could not be distinguished by \( J_{\text{link}} \) values. In all cases, top large communities have large \( J_{\text{link}} \) values; for instance, in the case of OLED, there exist dense linkages among the top three communities, #1 multilayer devices (7,032 papers, published in 2004.1 on average), #2 luminescent property (6,094 papers, published in 2004.4 on average) and #3 polymer (4,128 papers, published in 2003.3 on average). Since the size of these three communities is much larger than the rest, as shown in Table 2 (2), \( J_{\text{link}} \) values were affected by the community sizes.

In Figure 2B-1, B-2, B-3, compared to the related communities, the unrelated communities tend to have large \( L \) values. We can draw a horizontal line to distinguish, although there are a few errors. When we define the boundary of \( L \) as:

\[
L_{\text{boundary}} = \max_{c \in C_{\text{related}}} \{ L(c) \}
\]

where \( C_{\text{related}} \) is the sets of communities our experts judged as related to the topics, \( L_{\text{boundary}} \) values become 5.27 (\( L(10) \)), 2.84 (\( L(9) \)) and 3.99 (\( L(5) \)) in (1)NBT, (2)OLED, and (3)INV, respectively. Then there were three errors in total; #12 nanobacteria (132 papers, published in 2005.9 on average) and #15 nanosecond pulls (111 papers, in 2005.5) for (1)NBT, #11 magnetoresistance in organic semiconductors (59 papers, in 2005.9) for (2)OLED and no error for (3)INV. \( L \) seems to show better performance except in the above few cases. The reason why we cannot distinguish these cases as an unrelated community is discussed later.

In Figure 2C-1, C-2, C-3, the unrelated communities could not be distinguished by \( J \). The values of \( J \) were close to 0.2 and slightly decreased as the rank becomes larger but did not vary widely regardless of the sizes of communities. In Figure 2D-1, D-2, D-3, it was also difficult to judge which communities were unrelated by the values of \( C \). The values of \( C \) decreased as the community sizes got smaller. For top communities, \( C \) values were close to 1.0, which meant that almost all phrases had similar \( tf-idf \) weights.

Discussion

We compared four measures of relatedness between each pair of communities. As a result, the method measuring the distance from the topological center in a citation network, \( L \), is the best way of detecting the unrelated communities because the unrelated communities tend to be far away from the center in a citation network.

The interpretation of each measure is shown in Figure 3. In this figure a thick line represents a close relationship. Regarding \( J_{\text{link}} \), the smaller the rank is, the larger the \( J_{\text{link}} \) is and the smaller the \( \arg\max\{J_{\text{link}}\} \) is, which means core and large communities tend to be tightly connected to each other. On the contrary, small communities tend to be weakly connected to each other but connected to larger ones more tightly, as where the rank is large, \( \arg\max\{J_{\text{link}}\} \) is small. As the size of
<table>
<thead>
<tr>
<th>id</th>
<th>#Papers</th>
<th>Average publication year</th>
<th>Topics</th>
<th>Unrelated by experts</th>
<th>$J_{link}$</th>
<th>$\arg\max{J_{link}}$</th>
<th>$L$</th>
<th>$J$</th>
<th>$\arg\max{J}$</th>
<th>$C$</th>
<th>$\arg\max{C}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>1</td>
<td>9113 2006.5</td>
<td>Nanostructures</td>
<td>5.75%</td>
<td>2</td>
<td>2.97 0.20</td>
<td>2</td>
<td>0.96</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7855 2006.8</td>
<td>Bio-imaging</td>
<td>5.75%</td>
<td>1</td>
<td>2.46 0.20</td>
<td>1</td>
<td>0.96</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6088 2006.1</td>
<td>Drug delivery and biomedical applications</td>
<td>4.42%</td>
<td>2</td>
<td>3.55 0.20</td>
<td>2</td>
<td>0.96</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3361 2006.8</td>
<td>Carbon nanotube and biosensors</td>
<td>5.02%</td>
<td>2</td>
<td>2.81 0.20</td>
<td>5</td>
<td>0.92</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2806 2005.8</td>
<td>MEMS</td>
<td>2.76%</td>
<td>1</td>
<td>3.58 0.20</td>
<td>4</td>
<td>0.95</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>868 2000.8</td>
<td>Nano-plankton</td>
<td>0.05%</td>
<td>5</td>
<td>7.52 0.20</td>
<td>7</td>
<td>0.79</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>636 2005.2</td>
<td>Mass spectrometry</td>
<td>0.29%</td>
<td>5</td>
<td>5.20 0.20</td>
<td>6</td>
<td>0.86</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>260 2005.0</td>
<td>Magnetic fluids</td>
<td>0.28%</td>
<td>12</td>
<td>4.08 0.19</td>
<td>9</td>
<td>0.73</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>258 2006.5</td>
<td>Nano-silicon</td>
<td>0.35%</td>
<td>5</td>
<td>3.60 0.19</td>
<td>8</td>
<td>0.77</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>247 2005.6</td>
<td>Membrane and filtration</td>
<td>0.07%</td>
<td>1</td>
<td>5.27 0.19</td>
<td>9</td>
<td>0.71</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>234 2006.3</td>
<td>Biological tissues</td>
<td>0.28%</td>
<td>5</td>
<td>4.53 0.19</td>
<td>9</td>
<td>0.73</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>132 2005.9</td>
<td>Nano-bacteria</td>
<td>0.28%</td>
<td>8</td>
<td>4.01 0.19</td>
<td>14</td>
<td>0.64</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>113 2004.5</td>
<td>Nano-colloid</td>
<td>0.01%</td>
<td>3</td>
<td>7.24 0.18</td>
<td>12</td>
<td>0.56</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>111 2002.8</td>
<td>Measurement</td>
<td>0.01%</td>
<td>4</td>
<td>6.14 0.19</td>
<td>12</td>
<td>0.62</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>68 2005.5</td>
<td>Nanosecond pulls</td>
<td>0.05%</td>
<td>5</td>
<td>4.90 0.15</td>
<td>12</td>
<td>0.46</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>1</td>
<td>7032 2004.1</td>
<td>Multilayer devices</td>
<td>12.93%</td>
<td>2</td>
<td>1.93 0.23</td>
<td>2</td>
<td>0.97</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6094 2004.4</td>
<td>Luminescent property</td>
<td>13.01%</td>
<td>3</td>
<td>1.73 0.23</td>
<td>1</td>
<td>0.97</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4128 2003.3</td>
<td>Polymer</td>
<td>13.01%</td>
<td>2</td>
<td>1.80 0.22</td>
<td>2</td>
<td>0.97</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>673 2003.9</td>
<td>Polymerization</td>
<td>0.10%</td>
<td>3</td>
<td>5.25 0.19</td>
<td>5</td>
<td>0.89</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>277 2003.2</td>
<td>Polymerization</td>
<td>0.08%</td>
<td>4</td>
<td>7.86 0.19</td>
<td>6</td>
<td>0.80</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>263 2005.9</td>
<td>Light curing</td>
<td>0.00%</td>
<td>1</td>
<td>4.64 0.19</td>
<td>5</td>
<td>0.63</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>135 2003.9</td>
<td>Temperature dependence of light emission</td>
<td>0.35%</td>
<td>1</td>
<td>2.02 0.20</td>
<td>8</td>
<td>0.70</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>91 2002.5</td>
<td>Carbazole for OLED</td>
<td>0.32%</td>
<td>2</td>
<td>2.14 0.20</td>
<td>7</td>
<td>0.66</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>83 2006.5</td>
<td>Thin film transistors for OLED</td>
<td>0.07%</td>
<td>1</td>
<td>2.84 0.19</td>
<td>10</td>
<td>0.58</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>82 2004.7</td>
<td>Luminescent metal-organic frameworks</td>
<td>0.02%</td>
<td>1</td>
<td>4.45 0.20</td>
<td>7</td>
<td>0.62</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>59 2005.9</td>
<td>Magnetoresistance in organic semiconductors</td>
<td>0.19%</td>
<td>1</td>
<td>2.15 0.19</td>
<td>7</td>
<td>0.58</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>47 2002.5</td>
<td>Columnar discotics for OLED</td>
<td>0.17%</td>
<td>2</td>
<td>1.77 0.18</td>
<td>8</td>
<td>0.57</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>42 2004.6</td>
<td>Nanostructures for inorganic LED</td>
<td>0.05%</td>
<td>3</td>
<td>3.86 0.17</td>
<td>11</td>
<td>0.41</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>28 2002.4</td>
<td>Polymerization</td>
<td>0.00%</td>
<td>1</td>
<td>6.21 0.15</td>
<td>12</td>
<td>0.40</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>21 2002.0</td>
<td>Polymerization</td>
<td>0.01%</td>
<td>2</td>
<td>3.33 0.14</td>
<td>12</td>
<td>0.38</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
community drops significantly at a certain rank and the $J_{link}$ follows this trend, $J_{link}$ can be considered affected by the size of community significantly.

It is difficult to distinguish unrelated communities by semantic information like $J$ and $C$, as shown in Figure 3C,D, as most communities share many common terms. For instance, most papers in NNB contained nano and bio and these affected the semantic relatedness even if we applied tf-idf. Regarding $J$ values, the more similar the sizes of communities are, the larger the $J$ values are. Moreover, communities share many terms so that the Jaccard coefficients are almost the same values around 0.2 for all communities. $C$ values also fail to distinguish unrelated communities because the terms are so overlapped that it is difficult to discover the differences among communities. Larger communities tend to have large coverage of terms in the targeted research domain.

However, when we measure the relatedness by $L$, the unrelated communities tend to have larger $L$ values than the related communities, as shown in Figure 3B. The unrelated communities are far away from the geographic center in the citation network. Although there exist a few errors, this measure can be regarded as the best one among the four we compared. As described above, there were three communities with larger $L$ values than the boundary. A detailed analysis of why these three exceptional communities could not be classified correctly is as follows.

**NBT #12, Nanobacteria**

In Community #12 (nanobacteria), the main issue is to characterize nanobacteria. For example, the most cited paper in that community (Khullar et al., 2004) extracted the nanobacteria from human renal stones from a north Indian population and characterized its structure by using microscopy and another analyzing technique. Another paper (Sommer et al., 2003) investigated the durability of nanobacteria under environmental stress situations and discussed the process of life formation in a primordial soup, or other early wet environment. These topics differ from the issues in the mainstream of nanobiotechnology such as drug delivery and electric devices utilizing nanotechnology. However, one paper (Kim et al., 2004), which is not included in Community #12 but is the most cited paper, receiving nine citations from papers in Community #12, investigated electrochemical activity of microbial consortium and found that it is enriched by utilizing fuel cells with organic wastewater as the fuel. The existence of this interdisciplinary paper is the reason why our approach cannot distinguish Community #12 as digressional from the mainstream of nanobiotechnology.
FIG. 3. The interpretations of each measure. (A–D) The link-based Jaccard coefficient, $J_{\text{link}}$, the average shortest path from the center of the graph, $L$, Jaccard coefficient of term vectors, $J$, and cosine similarity of tf-idf vectors, $C$, respectively.

NBT #15, Nanosecond Pulls

In Community #15 (nanosecond pulls), the main issue is to apply nanopulse lasers to characterize biological materials whose size is not always on the nanometer order and therefore is not the focus of mainstream nanobiotechnology, where the size of biological materials is on the order ranging from hundreds to several nanometers. In Community #15 the term nano is used to represent time duration but not size. However, the papers in this community sometimes cite the papers in the other communities that utilize nanopulse lasers to characterize or manufacture nanobio materials. This overlap and term ambiguity (nano has meanings of both time and scale) is the reason why our approach cannot distinguish Community #12 from the other mainstream communities.

OLED #11 Magnetoresistance in Organic Semiconductors

In Community #11 (magnetoresistance in organic semiconductors), the main issue is the effect of giant magnetoresistance of organic materials but not electroluminescence. Although they studied different physical properties of organics, i.e., magnetoresistance and electroluminescence, there is a common mechanism, spintronics, for both properties, and spintronics is regarded as the key to understanding the multifunctionality of organic materials including electronic, optical, and magnetic properties for energy conversion, optical communication, and sensing technologies (Hu, Yan, & Shao, 2009). Actually, a paper in Community #11 (Davis & Bussmann, 2004) investigated the magnetic field effects to control spin polarized currents to manipulate electroluminescence in organic light-emitting diodes. Similarly, another paper in Community #11 (Mermer, Veeraraghavan, Francis, & Wohlgenannt, 2005) investigated the effect of the magnetic field on the electroluminescence of the devices and analyzed the relationship between magnetoresistance and electroluminescence. Therefore, our approach cannot distinguish Community #11 focusing on magnetoresistance from the other mainstream OLED communities because of the existence of hidden common topics, i.e., spintronics.

It is noteworthy to point out that we can observe the relationship between $J_{\text{link}}$ and $Q_{\text{max}}$. In Table 1, the values of $Q_{\text{max}}$ vary from 0.39 to 0.61. If the core and large communities are connected strongly each other, $Q_{\text{max}}$ will be small, as there are many intercommunity edges. On the other hand,
$J_{link}$ becomes larger because the link density between two communities is large. In the case of NBT, $Q_{max}$ was 0.61 and $J_{link}$ of top four communities were between 4% and 6%. On the contrary, In the case of OLED $Q_{max}$ was 0.39 and $J_{link}$ of top three communities were around 13%. This different characteristic among corpus might result in different accuracy to measure relatedness by link similarities, while we cannot observe different tendencies among the corpus we used.

These three failure cases show both limitations and possibilities. Our approach based on topological distance can, in most cases, clearly distinguish mainstream communities and digressional communities, and therefore can be used to build a relevant corpus from the initial corpus collected simply by the query and includes noise.

However, it does not work when interdisciplinary or common issues exist like the case of NBT and OLED. Some communities do not obey a general trend and cannot be distinguished, which is a limitation of our approach. The most interesting suggestion from the detailed analysis with experts is that all of the three communities focus on the interdisciplinary but fundamental technologies, which can be applied for wide purposes other than our targets. There is a possibility that we can extract hidden interdisciplinary and common issues by focusing on these exceptional cases that do not obey a general trend of topological distance measures.

Conclusion

As academic disciplines are segmented and specialized, it becomes more difficult to capture research areas precisely by either keywords or journal categories. The aim of this study was to propose a method of measuring the relatedness among sets of academic papers in order to detect unrelated communities from the targeted topic. A citation network, extracted by given keywords, is divided into communities based on the density of links. We measured and compared four measures of relatedness between two communities in a citation network with three large-scale citation datasets. Among the four, the distance from the center in a citation network is the best measure for removing the unrelated communities, while it is the prerequisite for using this measure that the centered node is relevant to the topic. The other three measures, the ratio of the number of intercluster links over the all links, the ratio of the number of common terms over all terms, cosine similarity of tf$\cdot$idf vectors, could not distinguish well because (1) intercluster links were dense among large clusters but sparse among in other cases, and (2) since most clusters share many common terms, it is difficult to distinguish unrelated communities by semantic information. We also found that although there are a few errors in distinguishing the unrelated communities, these communities focus on interdisciplinary and fundamental technologies. And as we can automatically detect these communities, our approach can be used to extract hidden common or interdisciplinary topics among a pile of academic papers.

Acknowledgments

One of the authors (N.S.) was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (Start-up), 21800011, 2009. Another author (Y.K.) was partially supported by the New Energy and Industrial Technology Development Organization (NEDO), Grant for Industrial Technology Research (09D47001a). Y.K. was also supported by the Ministry of Education, Science, Sports and Culture (MEXT), Grant-in-Aid for Young Scientists (B) (21700266).

References


