

Title	類似度分析を用いた科学技術知識のマッピング手法
Author(s)	梶川, 裕矢; 柴田, 尚樹; 坂田, 一郎
Citation	年次学術大会講演要旨集, 25: 822-824
Issue Date	2010-10-09
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/9419
Rights	本著作物は研究・技術計画学会の許可のもとに掲載するものです。This material is posted here with permission of the Japan Society for Science Policy and Research Management.
Description	一般講演要旨

類似度分析を用いた科学技術知識のマッピング手法

○梶川裕矢, 柴田尚樹, 坂田一郎 (東大)

要旨

科学技術の知の構造や発展を計量書誌学の立場から記述する試みが多くなされるようになってきたが、分析結果を左右する対象領域の設定の方法や、異なるデータベースの接続など、今後検討すべき研究課題も多く残されている。本研究発表では、テキストマイニングおよびリンクマイニングによる類似度分析を用いることで、上記の課題を克服できるだけでなく、サイエンスリンケージに代わる産学の関連性を測る新たな指標として活用できることを示す。

1. はじめに

計量書誌学ならびに、科学知識のマッピング手法は、科学技術イノベーション政策の科学の中の重要な要素である。計量書誌学の手法を用いることで、特定の分野での研究の動向を客観的で定量的な情報を用いて記述することができる。そのような手法の代表的なものとしては、各国、各研究機関、各研究者ごとに、出版論文数や被引用回数を集計し、ランキングを行う方法や、急速に被引用を集めている論文をモニタリングすることで、注目度の高い論文をホットペーパーとして提示するといった、統計的な集計量に基づくアプローチが既に各国で実施されている。

しかし、そのような単純な集計量に基づくアプローチでは、予め設定された粒度の分野でしか分析できないことや、論文や各研究領域間の関係性を分析できないという問題点があり、現在、テキストマイニングやリンクマイニングを用いた高度な分析手法の開発が東京大学他、欧米の各研究機関において進められている(梶川, 2010)。

例えば、サイエンスリンケージとは、特定領域において 1 特許あたり引用されている学術論文の数で、その領域において、基礎研究と産業応用の関連の程度を表す指標とされている。しかし、そのような分析では、論文と特許の双方が存在する領域における何らかの論文に代表される基礎研究と、特許で代表される産業応用の間の何らかの繋がりを表すことができるものの、基礎研究は存在するものの産業応用がなされていない領域や、特許として多くの出願があり産業上の重要性が高いにもかかわらず、基礎研究が不十分な領域など、学術と産業技術の間に想定される様々な関係性を十分に捉えきることができない(Shibata, Kajikawa & Sakata, 2010)。

そこで、本研究では論文と特許の関係性を例として、科学技術領域間の関係性を捉えるための指標の開発を行う。以下、本研究に用いたデータの手法の説明、分析結果の提示、応用事例の紹介の順に述べる。

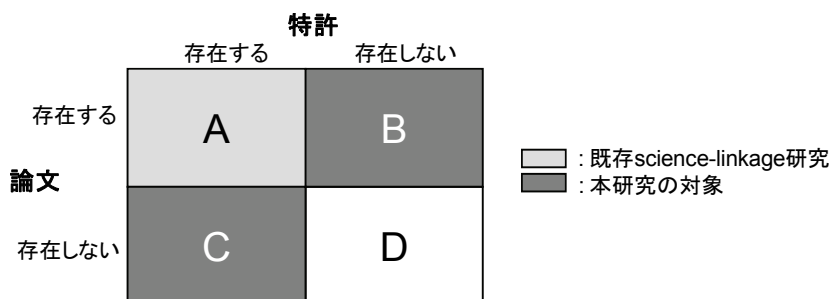


図1. 論文と特許の関係性

2. データと分析手法

本研究では、分野間の関係性を測定する3つの指標を提案し、それら指標の有効性を、太陽電池における論文と特許の類似性の測定を用いて検討する。分析の手順を図2に示す。

分析ではまず、”solar cell*”を検索語として、特許情報、論文情報を収集した(1)。特許のデータベースは USPTO、論文データベースはトムソンロイター社の Science Citation Index を用いた。得られたデータを用いて、特許、論文をそれぞれノードとし、直接引用をリンクとする特許引用ネットワーク、論文引用ネットワークを作成した(2)。得られたネットワークのうち、最大連結成分を分析の対象として設定し、それを抽出した(3)。最大連結成分に対し、Newman 法を用いたクラスタリングを行い、クラスタに分割した(4)。最後に、各特許クラスタと各論文クラスタの意味的な類似度を測定した。類似度の測定には、Jaccard 係数(d_j)、tf-idf の Cosine 類似度(d_c)、log 重み付きの tf-idf の Cosine 類似度を用いた(d_{lc})。

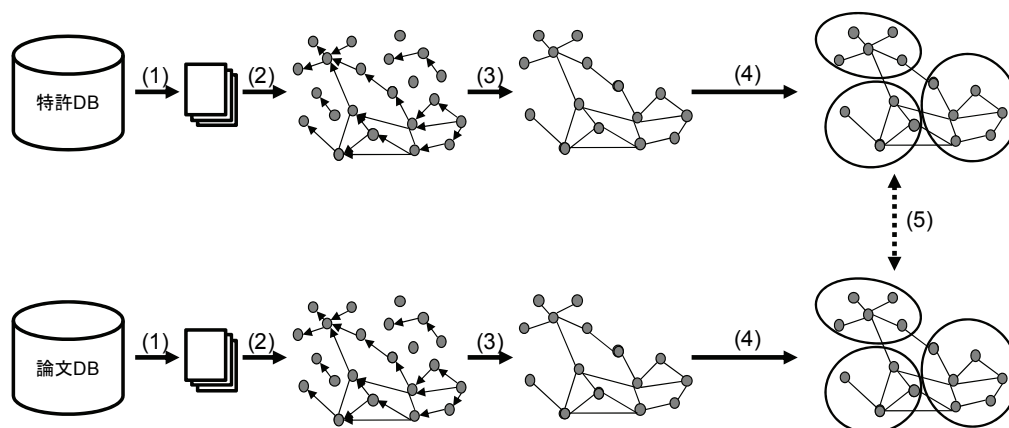


図2. 分析の手順. (1) データ取得、(2) 引用ネットワーク作成、(3)最大連結成分抽出、(4) クラスタリング、(5)類似度測定。

3. 結果

本研究では、特許クラスタ、論文クラスタともサイズの大きいものから順に上位 10 クラスタを分析の対象とした。太陽電池の特許クラスタ、論文クラスタを手動で分析したところ、両者ともにシリコン系太陽電池が最大であり、共通に現れることが分かった。また、化合物系太陽電池も論文クラスタ、特許クラスタともに現れており、論文クラスタ中では 2 番目に大きく、特許クラスタでは 6 番目に大きい領域であった。論文クラスタ、特許クラスタを各クラスタに含まれるノード数で測った順位をそれぞれ s_r , t_r とすると、 d_j 、 d_c 、 d_{lc} のうち、 $(s_r, t_r)=(1,1), (2,6)$ に対する類似度が高いと評価できる指標が、論文と特許の意味的な距離を測定する上で有効性の高い指標であると評価できる。

表1は、太陽電池を事例として、3 指標の評価を行った結果である。表1からは d_c が最も類似性の測定指標として有効であることが分かる。例えば、 $(s_r, t_r)=(1,1)$ において、 d_c は最大値 3.464 をとる。これは類似度の高い科学技術領域間を類似度が高いと適切に評価できていることを示している。

4. 応用事例

本研究で提案した指標のうち、最も良好な結果を得た d_c を用いて二次電池分野の分析を行った。これを図1のフレームワークを用いて表現した結果が図3である。

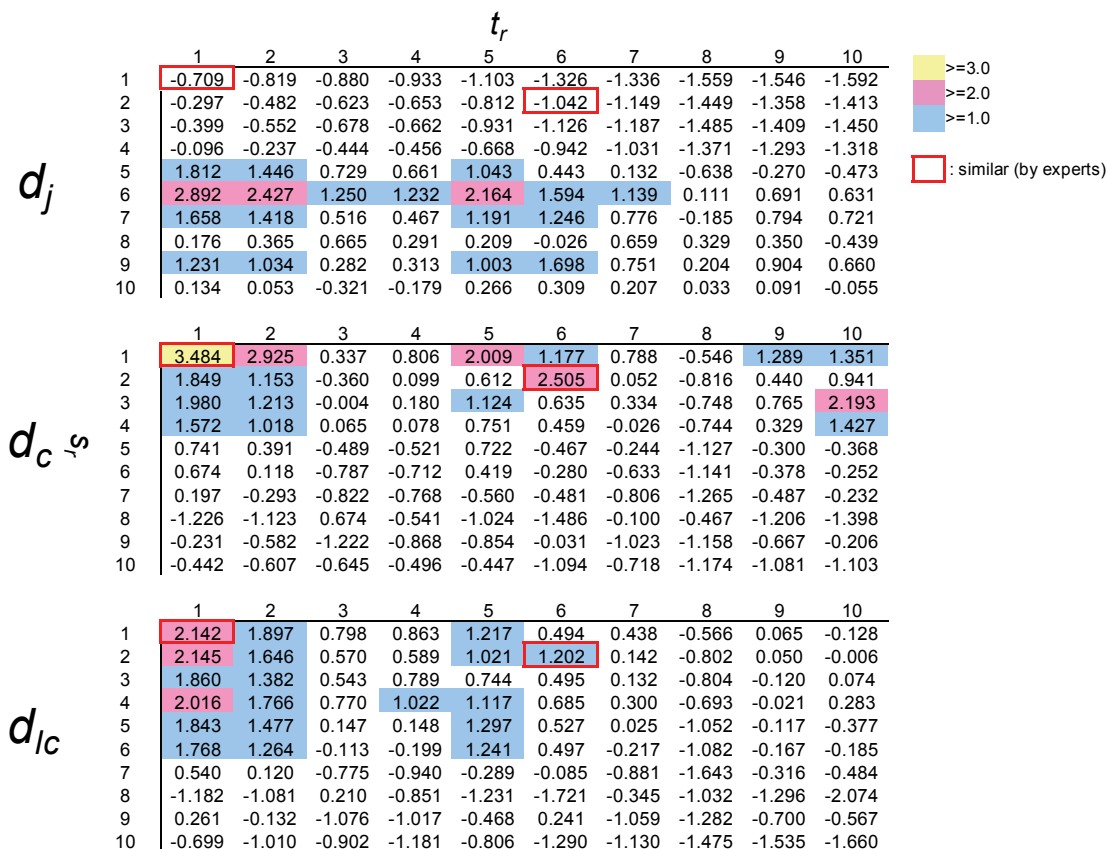
図3に示している通り、Li イオン電池、セルデザイン、メタルハイドライドといった領域は、論文、特許いずれにおいても主要な領域であることが分かった。一方、自動車への応用は特許クラスタのみに現れており、論文クラスタとしては存在しない。一方、固体電解質に関するクラスタは論文としては主流であるものの、主要な特許クラスタには存在しなかった。これは、液漏れ防止技術として、論文では固体電解質がメインだが、特許では機械的なパッケージングがメインであることに由来するものと思われる。

従って、本研究で提案した指標を用いて、論文と特許の関係性の分析を行い、学術研究としては多く行われているものの、特許出願はあまり積極的になされておらず、今後、産業上の応用として重要性が高まるポテンシャルが高い科学技術領域を抽出できる可能性がある。日経エレクトロニクス(2010 年 5 月)では、東工大・菅野了次教授による「固体の材料だけで電池として機能するという認識がやっと出てきた」という談話や、2010 年 3 月の日本の電気化学会でトヨタが大学との共同研究の成果を中心に 7 本の発表を行ったという記事を発表しており、固体電解質が今後産業上重要であるという今回の分析結果は妥当なものと思われ

る。

今後の展開として、上記のような応用だけでなく、分野横断領域における各領域の関係性、分析の対象とするデータの内容の一致度の分析によるデータクレンジング等、様々な応用に対し、提案指標を用いた分析を行うことが考えられる。

図2 各指標の評価結果



		特許	
		存在する	存在しない
論文	存在する	Liイオン電池 セルデザイン メタルハイドライド	固体電解質*
	存在しない	自動車への応用	

図3 二次電池分野における分析結果

参考文献

N. Shibata and Y. Kajikawa and I. Sakata, "Extracting the commercialization gap between science and technology - Case study of a solar cell -", *Technological Forecasting and Social Change*, 77 (2010) 1147-1155.

梶川裕矢, 「リンクマイニングを用いた引用情報の活用」、*情報の科学と技術* 60 (2010) 224-229.

日経エレクトロニクス『次世代電池は「全固体」へ』2010年5月