

“テキストマイニングを活用した技術動向分析”

～英語論文の分析のご紹介～

有門 経敏、コーポレート開発本部、東京エレクトロン(株)

1. はじめに

読者の方々は、半導体集積回路をご存じのことと思います。皆様お使いのパソコンやスマホの中にギッシリ詰まっている黒い四角いもの、それが半導体集積回路です。演算機能を持つ MPU (Micro Processing Unit)、情報を記憶する DRAM (Dynamic Random Access Memory) など各種半導体集積回路の働きによって私たちの快適な “Electronic Life” が実現されています。かつて “産業の米” と呼ばれた半導体集積回路は今や “社会の米” と呼ばれるほど社会の隅々にまで深く浸透し、私たちの生活に必要不可欠なものになりました。

東京エレクトロンは半導体集積回路を製造する装置の開発、製造、販売およびサービスを行う会社です。筆者はコーポレート開発本部にて新製造技術の探査、ならびに開発企画を担当しています。

半導体産業は時代の最先端を行くハイテク産業のため、熾烈な研究開発競争を繰り広げています。研究開発費の売上高比率は 10% を越え、あらゆる産業の中で最も高い部類と思われます。ただし、研究開発費が多ければ良いというものではなく、そこに効率が求められることはいうまでもありません。

企業の研究開発も時代と共に様変わりしました。昔、大手企業は中央研究所を整備し、新しい技術を生み出すことに力を注ぎました。ところが全て自前でやっている競争に間に合わない。必要な技術は社内で生み出すのではなく外部から調達するものという考え方が広がってきました (参考文献 西村吉雄著 「産学連携—中央研究所の時代を超えて」)。つまり、自社のビジネスに必要な技術シーズを社外から調達し、社内では量産技術に仕上げることに注力するのです。この方が、一から社内で開発するよりも早いからです。

さて、技術を調達するためには世界に散在する研究成果や技術シーズを探し、個々の技術を評価しなければなりません。一つの技術を評価する際に必要な作業は、(i) その技術に関わる世界の動向を描き出して技術マップを作成すること、(ii) そのマップ上に当該技術をプロットすることです。そして、当該技術を開発すべきか否か、開発するならば、どこからシーズを調達するのか、などなどを判断致します。

この作業において収集する情報量は多く、手作業で分析するのは極めて困難です。そこで、筆者は情報分析のツールとしてテキストマイニングを使っています。本講演

では2つの事例を取り上げて、筆者の行っている分析プロセスを紹介します。

2. データベースを用いる情報収集

私たち開発企画部門が日常的に取り扱う技術情報は特許と論文です。最近ではデータベースが整っていますので、特許や論文調査が随分容易になりました。ただし、世界中の特許と論文が調査対象ですので、言語はいうまでもなく英語です。

特許のデータベースでは通常、出願人、出願日、抄録、請求項などが項目毎にエクセルに出力されますので、そのままテキストマイニングを実行できます。

一方論文データベースの場合、関連論文をリストアップできますが、タイトルや abstract をひとまとめにしてエクセルに出力してくれません。そこで、筆者は通常、abstract をコピーしてエクセルにペーストすることによってテキストマイニング用にデータを作成しています。

3. 事例—Electrochromic Display の技術動向分析

最初に、Electrochromic Display の技術動向について調査した結果を説明します。Electrochromic Display は電子ペーパー用 Display の一つで、電子書籍リーダーやデジタルサイネージなどの応用を目指しています。データベースを使って過去 10 年間の関連論文を 70 件抽出しました。

まず論文数の年次推移をグラフにしたところ、2003 年以降増加していますが、2006 年以降は飽和傾向に転じたことがわかりました。開発する企業数は限られており、実用化されてもニッチ市場に止まるものと推察されます。

次に内容分析について述べます。Abstract と発行年をコピーしてエクセルに貼り付け、分析データを作成しました。

表 1 2005 年～2008 年の出現頻度上位 20 位の一覧

	2005年		2006年		2007年		2008年	
1	redox	11	polymer	10	display	14	electrochromic	14
2	film	9	electrochromic	9	electrochromic	12	device	12
3	bpy	7	display	6	film	12	display	12
4	electrochromic	7	film	6	cell	11	viologen	12
5	device	5	electrode	5	have	9	have	8
6	display	4	material	5	ITO	7	polymer	8
7	electroluminescence	4	potential	5	PANI	7	electrode	6
8	response	4	application	4	device	6	show	6
9	two	4	aqueous	4	process	6	time	6
10	electrode	3	center	4	removal	6	use	6
11	mode	3	electrochemical	4	efficiency	5	electrochemical	5
12	nanotube	3	response	4	electrochemical	5	microspheres	5
13	polymer	3	time	4	laser	5	system	5
14	ruthenium	3	absorption	3	microspheres	5	ZnO	5
15	SCV	3	base	3	nanofibre	4	area	4
16	use	3	dot	3	property	4	array	4
17	WRP/RP	3	electrolyte	3	solar	4	cell	4
18		0.14	investigate	3	substrate	4	electrolyte	4
19		0.5	ITO	3	surface	4	good	4
20		0.61	oxidation	3	thin	4	molecule	4

TMS を用いて上記のデータを 1 年ごとに分かち書きした後、単語出現頻度を出力

します。グリッド表示にすると出現した単語全てを見ることができます。

ここで TMS の裏技の登場です。グリッド表示の左上隅に何も書いていない小さなボタンがあります。このボタンを右クリックすると望みの領域を選んでコピーし、エクセルにペーストすることができます。このコピー&ペーストを繰り返すことによって 2005 年から 2014 年分までの単語出現頻度の一覧表を作成しました。一覧表全体をこの紙面に貼り付けると字が小さくて読めませんので、2005～2008 年の 4 年分の上位 20 位のみ表 1 に示します。

単語には毎年高い頻度で出現するものと特定年のみに出現する単語とがあります。出現頻度は技術動向を反映しているものと考えて、以下に述べるやや手の込んだ処理をしました。

“Display”という単語を例にとって処理プロセスを説明しましょう。表 2 に Display の出現頻度と累計頻度とをまとめました。上の欄にそれぞれの年の出現頻度、下欄に累計の頻度を示しています。累計頻度は次のように計算します。例えば 2005 年と 2006 年の出現頻度はそれぞれ、4 回と 6 回です。そこで、4 回と 6 回を足して 2006 年の累計出現頻度を 10 回としました。また、2007 年は出現頻度が 14 回なので 2006 年までの累計頻度 10 回に 14 回を加えて 24 回と算出しました。このような計算を各単語について繰り返すことによって、それぞれの単語の累計頻度を計算しました。

表 2 単語“Display”の出現頻度

	2005	2006	2007	2008	2009	2010	2011	2012	2013
出現頻度 (単年度)	4	6	14	12	12	12	17	8	12
出現頻度 (累計)	4	10	24	36	48	60	77	85	97

次に、各単語に年度内で累計発生頻度の多い順に順位を付けます。この作業によって累計出現頻度の高い順に単語が並んだ表ができあがります。この順位を年に対してプロットしたのが図 1 です。

この解析では毎年出現する単語は継続的に上位を占めますが、出現頻度の減少した単語の順位は下がります。この“上がったりがったりの傾向”から動向を読み取ることができます。

例えば、“Redox”は Electrochromism(■)の基本メカニズム関わる単語ですが、年々順位が低下しています。このことから基礎研究は 2005 年頃終了し、その後は応用研究に軸足が移ったと推察できます。また、Microsphere という単語を見てみましょう。Electrochromic 材料と電極との接触抵抗を下げるために提案されたアイデアですが、2007 と 2008 年に上位に進みましたが、その後順位は低下し 2011 年には欄外に

消えました。このことは、この技術が接触抵抗を下げる決め手にはならなかったことを意味しています。

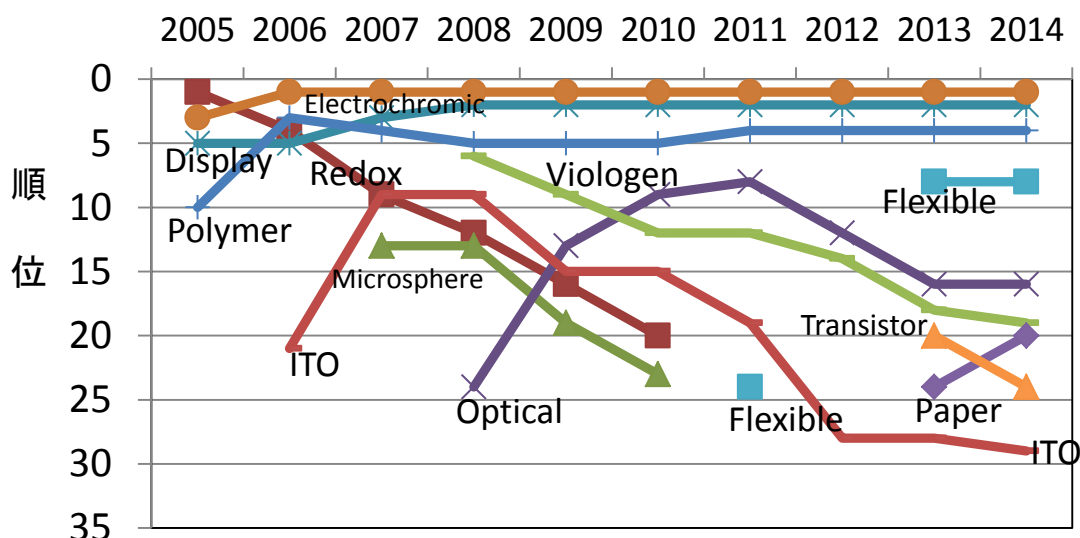


図1 主要な単語の順位の変化

このようにして一つ一つの単語について解釈していきます。本発表の主旨はテキストマイニングの使い方の提案であって Electrochromic Display の技術動向の説明ではありませんので、これ以上の詳細な解釈は割愛させていただきますが、出現頻度の推移だけで技術動向をある程度描き出せることがお分かりいただけたことと思います。TMS は共起関係や係り受けについても同様に、出現頻度の年次変化を分析する機能を備えています。それらの機能を活用すれば、さらに詳細な分析ができることは言うまでもありません。

4. まとめ

グローバル化時代といわれる現代にあつては、市場や競争相手が海外にすることが少なくありません。技術開発においても競争相手は海外の研究機関、大学、および企業です。

特許や技術論文などの技術情報を得るにはデータベースが便利です。海外の特許や論文動向を知るためにはやはり外国製のデータベースが、カバー率の点で優れているようです。使用されている言語が英語であることはいうまでもありません。

世界の特許や論文を集めるとその数は膨大な量に上り、一つ一つを読んで理解することはほとんど不可能です。

そこで筆者はテキストマイニングを活用しています。TMS は筆者に替わって文書を読み、筆者が読まねばならない文章だけを抽出するツールで、筆者のやるべき作業

を大幅に軽減してくれるアシスタントです。

本報告では単語の出現頻度から技術動向を描き出した事例を紹介しました。

使ってみて感じることは、上手に使うためには使う側に技量が必要なことです。自分の能力を開発すればするほど TMS を上手く活用できると思います。筆者にとって TMS は、共に進化する仕事のパートナーといえます。