ＡＩを用いた無効資料調査について

弁理士　鈴木　守

現在、ＡＩを利用した特許調査ツールが多く存在する。将来は、ＡＩを使った調査が可能になるというような話もあり、その流れに乗り遅れまいとＡＩ検索ツールを導入し、果たして調査が可能かどうかを検討してみた。

なお、本稿は Patent field 株式会社が提供しているＡＩ特許総合検索サービス（以下、「Patentfield」という。）の使用経験に基づくものであり、必ずしも現存するすべてのＡＩ調査ツールに当てはまるものではないが、大きくは外れていないと思う。

はじめに

（１）Patentfield のＡＩ検索機能

Patentfield は、公報番号や自由に発明を記述した文章に基づいて特許検索を行うセマンティック検索と呼ばれる機能を有している。公報番号を用いることと文章を用いることは異なるようだがロジックは同じである。というのは、公報番号を入力して検索を行う場合にも、結局、公報中の文章をクエリとするからである。Patentfield では公報番号で検索を行う場合、当該公報に記載の文章の中で、「タイトル／要約／請求の範囲／明細書／審査官キーワード」を用いるか、「請求の範囲」を用いるか、「要約」を用いるか等を選ぶことができる。

セマンティック検索では、クエリである文章をベクトル化（特徴語とその重みに分解）したうえで類似のベクトルを有する公報を検索する。セマンティック検索は同じベクトルを持っている公報を見つけるのではなく、クエリのベクトルに近いベクトルを持つ特許公報を近い順に並べる。Patentfield の場合には、クエリの公報または文章に近い方から 10000 件（デフォルト設定）の特許公報が順位付けして表示される。

（２）理論的な限界

セマンティック検索を行う際には文章同士の類似度を計算することができるように、最初にクエリとなる文章と検索対象の文章をベクトル化する。このベクトル化において文脈が失われてしまうので、セマンティック検索では文脈を考慮した比較を行うことはできない。

一例として、「基板の上にアルミからなる層を備え、さらにその上に銅からなる層を備える」という構造を有する発明を探したいとする。この検索クエリは、例えば「アルミ:9.7821 銅:7.798154 層:5.548257 基板 :5.291978 上 :2.933625 備 え る :2.514469 備え:1.959087 なる:1.91536 その:1.602236」等というように特徴語とその重みに分解される。

上記のベクトルを元に検索を行うと、アルミ、銅、層、基板といった語が特徴語として含まれた公報との類似度が高いと判断されるが、その構造が基板－アルミ－銅という順番で積層されているかどうかまでは判断することはできない。

セマンティック検索において類似度が高い公報は、特徴的な文言の現れ方が似ている公報であるといえる。特徴的な文言の現れ方が類似していれば発明の内容も類似している可能性は高いと言えるが、検索結果の類似度は発明どうしの類似度を表すものではない。

テスト１（公報番号での検索）

セマンティック検索で所望の公報が見つかるかを実験した。実験方法は次のとおりである。

［テストデータ］

異議申立事件において特許が取り消された事件、無効審判事件で特許が無効にされた事件をピックア

ップし、その主引例の公報を正解とした。異議申立や無効審判で特許が取消または無効の原因となった公報は、異議申立や無効審判の対象特許にかなり近いと考えられるからである。

［検索条件］

クエリとして以下の３つのパターンを用いた。

①「タイトル／要約／請求の範囲／明細書／審査官キーワード」

②「タイトル／要約／請求の範囲／審査官キーワード」

③「要約」

［実験結果］

表１に実験結果を示す。順位は、何番目に近い公報として抽出されたかを示している。

「なし」は 10000 位以内に入っていなかったことを示している。

【表１】実験結果

	対象特許	正解の公報	①全文	②請求項	③要約
1	特許第 6694402 号 (異議 2020-700862) 飲食店の運営管理システム	特開 2013-149185	4 位	なし	なし
2	特許第 6573205 号 (異議 2020-700164) 一のユーザによるアプリケーションプログラムの利用に関する予測データを計算する情報処理装置、情報処理方法、プログラム及びマーケティング情報処理装置	特開 2017-176690	なし	なし	なし
3	特許第 6381715 号 (異議 2019-700161) 提供装置、提供方法、提供プログラム、決定装置、決定方法、及び決定プログラム	特開 2016-75490	なし	なし	なし
4	特許第 6034481 号 (異議 2017-700549) 広告配信システム及び方法、並びにプログラム	特表 2015-504225	なし	なし	486 位
5	特許第 5921690 号 (異議 2016-701081) ネットワーク上における添削指導サービス提供方法及びこれに用いられるウェブサーバ	特開 2003-345232	なし	なし	なし
6	特許第 6165455 号 (無効 2020-800106) 商品管理システム	特開 2000-48262	なし	なし	3816 位
7	特許第 6407464 号 (無効 2020-800073) 情報処理装置、情報処理方法、情報処理プログラム、端末装置およびその制御方法と制御プログラム	特開 2004-195006	なし	なし	なし
8	特許第 4443244 号 (無効 2013-800023) 著作物利用実績報告書作成システム	特開平 11-238088	11 位	30 位	なし

特許第 6694402 号について見ると、①全文をクエリとしたときに 4 位に見つかったが、②請求項、③要約をクエリとする圏外になってしまった。正解公報の全文の雰囲気が対象特許と似ていたから①全文では上位に見つかったが、②請求項や③要約で見ると違うのであろう。何をクエリにするかで、4 位と圏外というように結果は大きく違ってくる。

他の結果を見てみる。特許第 4443244 号も①全文をクエリとした場合に 11 位、②請求項をクエリとした場合に 30 位という好成績であった。しかし、それ以外の特許では、圏外か見つかったとしてもかなり下の方であった。

このようにクエリの違いによって結果が大きく異なる。無効調査の観点でいうと最大の問題点はどのようなクエリが正解かが分からないことである。セマンティック検索では、特許公報の番号を指定するほか自由にクエリを記述することが可能であるが、自由記述したクエリが正解かどうか分からない。検索結果として 10000 位まで順に並べて表示されても、検索結果をどこまで信用してよいかわからず、どこまで見ればよいのかさっぱり分からない。所望の公報は圏外かもしれないし、３０位くらいにあるのかもしれない。このような状態で上から公報を探していく気はしない。

テスト２（フィルタリング）

Patentfield ではセマンティック検索の際のオプションとして、文書中に出現するキーワードで検索範囲を絞り込む機能がある。文書中に出現するキーワードを指定することで検索範囲を限定することでノイズを少なくする機能である。

表 2 及び表 3 は、特許第 6034481 号と特許第 6165455 号の検索において、フィルタリングを行った結果を示す図である。

【表 2 】特許第 6034481 号の検索においてフィルタリングした例

	対象特許	正解の公報	①全文	②請求項	③要約
4	特許第 6034481 号 (異議 2017-700549) 広告配信システム及び方法、並びにプログラム	特開 2015-504225	なし	なし	486 位
	「広告」でフィルタリング		2419 位	1347 位	454 位
	「広告主」でフィルタリング		1131 位	790 位	287 位

【表 3 】特許第 6165455 号の検索においてフィルタリングした例

	対象特許	正解の公報	①明細書	②請求項	③要約
6	特許第 6165455 号 商品管理システム	JPA2000-48262	なし	なし	3816 位
	「商品管理」でフィルタリング		1181 位（6028 文献中）	86 位（6028 文献中）	693 位（6028 文献中）

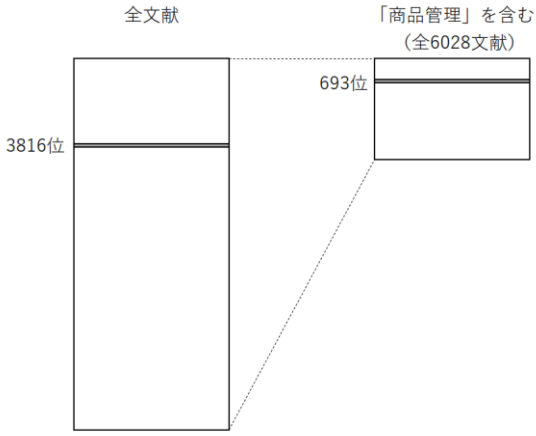
特許第 6034481 号において「広告」でフィルタリングを行うと、フィルタリングなしでは圏外だった①全文、②請求項を使った検索でも、10000 位以内に正解の公報が見つかるようになった。「広告」より狭い「広告主」でフィルタリングすると、さらに順位が上がった。

特許第 6165455 号については「商品管理」でフィルタリングすると母集団が 6028 文献になり、その中で所望の文献が検索された。母集団が 6028 文献というのは「商品管理」の文言を含む公報が 6028 文献ということであり、この 6028 文献とクエリの文章との近さを求めることになる。

このようにフィルタリングを行うと、所望の公報の順位は上がる。しかし、上がるといってもその効果は限定的で、いきなりトップにくることはない。考えてみれば、クエリは変わっていないのだから、これは当たり前のことである。

図１は、特許第 6165455 号について③要約でセマンティック検索した例を示す図である。フィルタリングを行っていない状態（図１左側）だと、全文献から③要約に近いベクトルを持つ文献を探してくる。この結果が 3816 位だった。これに対し、「商品管理」でフィルタリングすることにより母集団が減るが、フィルタリングをしなかったときに 3816 位より上位だった文献が全部なくなるわけではない（図１右側）。クエリが変わらない限り、元々上位だった文献より類似すると判断されることはなく、フィルタリングしただけではそれらを追い抜くことはできない。

【図１】



なお、「広告主」「商品管理」という文言はやや狭い文言であり、絞り込みのキーワードとしてどうなのか？という違和感を持たれた方もいらっしゃるかもしれない。無効資料調査においてこうしたキーワードで母集団を絞り込むことは難しい。種明かしをすると、今回は予め正解の公報が分かっていたので、恣意的に、正解の公報に含まれている文言を選択した。このため、所望の文献の順位をそれなりに上げることができた。実際には、フィルタリングによって順位を上げるのは容易ではないと思う。フィルタリングで順位を上げるためには、フィルタリングなしの場合に上位にあった文献を排除する必要があるが、そもそもそのような都合の良いフィルタがあるなら、ＡＩ検索を使うまでもないことになる。

まとめ

上に見たとおり、ＡＩによる検索ツールは所望の文献を一発で見つける可能性を秘めているがハズレも多い。たまたまトップ１０くらいに良いものが入っていればよいが、そうでない場合にどこまでその検索結果を見ていくべきかが皆目見当がつかない。

キーワード等による調査であれば検索式に狙いがあり、目論見と違う文献がたくさん入っていたりすれば、「ああこういう感じで引っかかっちゃうのか」と分かるので、検索式を修正することができるが、セマンティック検索の場合にはそうはいかない。

結論として、たまたま見つかればラッキーくらいの気持ちで使うことはできるが、セマンティック検索で無効資料を見つけるのは困難と思う。

なお、業者の方からは、Patentfield は無効資料調査を行うために最適化されたものではないという説明を受けた。理屈からしてそうだろうなとは思いつつ、実際のところを知りたくて試してみたが、無効資料調査の関係では難しいということが実感できた。ただし、予備的な調査や技術動向の調査には使える可能性は十分にあると思う。

最後になるが今回のテスト結果が再現するかどうかは分からないことにご留意いただきたい。なぜなら、教師データが増えることによりモデルのパラメータが変わるからである。出願日以前の公報だけを検索しているからそのようなことは起きないと思われるかもしれないが、これから出願される公報によって特徴語の重みが変わってくるので、類似度の計算結果が異なってくる。また、セマンティック検索のロジックに修正がなされる可能性もある。

KSI パートナーズ法律特許事務所
〒150-0031
東京都渋谷区桜丘町22-14 N.E.S.ビルS棟4階
TEL: 03-6455-3679
E-MAIL： patent@ksilawpat.jp
ksilawpat.jp