文書集合からのブール式自動生成

―概念検索と全文検索との橋渡し―

株式会社日立製作所　中央研究所／東京工業大学精密工学研究所　岩山　真

PROFILE　1992 年株式会社日立製作所入社。文書検索、自然言語処理等の研究に従事。また、NTCIR において特許検索用テストコレクションの作成に携わる。2009 年度より特許版産業日本語委員会委員。

makoto.iwayama.nw@hitachi.com　TEL 042-323-1111

1 はじめに

特許検索ではブール式を用いて全文検索を行うことが多いが、ブール式の構築にはノウハウや領域知識が必要となる。一方、概念検索は、思いついた文章を入力するだけで比較的精度の良い検索が行えるため、特に非専門家には有用な検索法である。反面で、検索基準がわかりにくい、制御が行いにくい、再現性（同じ質問文から同じ文書が数年後も上位に検索されるか？）に不安が残る、などの問題もある。

本稿では、任意の文書集合を指定すると、それのみが漏れなく検索できるブール式を逆生成する手法を紹介する。例えば、概念検索の結果が等価なブール式に変換できれば、概念検索から全文検索に移行できる。また、このブール式は概念検索の根拠とみなすこともできる。

以降、２節でブール式自動生成の事例を何点か示したあと、３節でブール式自動生成の手法を説明し、４節で手法の評価について述べる。

2 事例

（１）概念検索結果からのブール式生成

以下の請求項（特許電子図書館より引用）から概念検索を行ったとする。

> 頂面に燃焼室を凹設するとともに該燃焼室の周縁にスキッシュエリアを設けたピストンと、前記燃焼室に対向するドームを凹設したシリンダヘッドと、前記ドームの中央部に取り付けられて該ドームの側壁面に向って開口する噴口を備えた噴射ノズルを備え、前記ドームの外径を燃焼室の口径とほぼ同一に形成したことを特徴とするディーゼルエンジン。

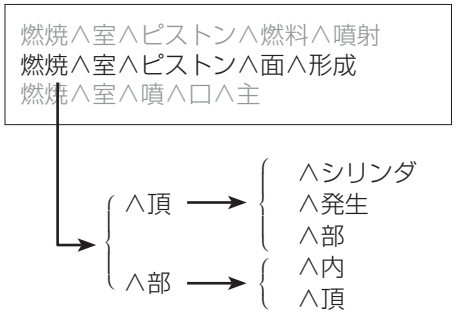
検索結果の上位 100 件から、３節で説明する提案手法を用いてブール式を逆生成してみる。

> 燃焼∧室∧ピストン∧燃料∧噴射∨
> 燃焼∧室∧ピストン∧面∧形成∨
> 燃焼∧室∧噴∧口∧主

この式で、指定した 100 件中、91 件が検索できる。ただし、全体で 494 件検索してしまうので、指定した 100 件のみが検索できる式にはなっていない。一般に、概念検索の上位文書のみを検索する式を生成することは難しい。それでも、生成された式を見ると、上位 100 件にどのような文書が含まれているか、おぼろげながら想像することができる。また、この式を使えば、指定した 100 件（実際は 91 件）がいつでも再現できる。

（２）絞り込み検索への展開

上記の式で、二番目の「燃焼∧室∧ピストン∧面∧形成」に興味を持ったとしよう。ユーザはこの部分式を使って、概念検索の結果を絞り込むことができる。さらに、この部分式の検索結果からパラメータを変えて長い式を生成させることもできる。



ここでは、論理積の長さが一つ長い式を生成させている。つまり、「燃焼∧室∧ピストン∧面∧形成」にヒットする文書集合は、「燃焼∧室∧ピストン∧面∧形成∧頂」と「燃焼∧室∧ピストン∧面∧形成∧部」の論理和にヒットする文書集合とほぼ等しいことがわかる。よってユーザは加わった新たなキーワードを頼りに絞り込みを続けていくことができる。これとは逆に、ヒット件数を増やすために論理積を短くすることもできる。この場合、指定した文書を含み、かつヒット件数が指定した文書数よりも多い検索式を生成することになる。

（3）異なる検索サービスとの連携

特許検索では、IPC などの特許分類を使えば、効率良く所望の文書を集めることができる。例えば、「H01F27」という IPC を使えば、「変圧器またはインダクタンスの細部一般」に関する特許のみを漏れなく集めることができる。IPC はキーワードや他の書誌情報と合わせて使うことも多い。

しかし、当然のことながら特許分類は特許にしか付与していないため、例えば、論文に対して同じような内容の検索を行うことはできない。ここでも、IPC で検索した検索結果から一度ブール式を逆生成すれば、論文検索においても、あたかも IPC を使っているかのような検索を行うことができる。上記の「H01F27」で特許を検索した結果から実際に式を生成させてみる。

```
コイル∧トランス∨
巻∧線∧絶縁∨
巻∧線∧次∨
目的∧変圧∧器
```

この検索式により「H01F27」からの検索結果が再現（ただし、精度や再現率は 100% でない場合が多い）できる。つまり、可能な限り「H01F27」と等価なブール式となっている。ほとんどの検索システムでは、キーワードからなるブール式を受け付けるため、上記の式をそのまま論文検索システムに入力すれば、近似的に「H01F27」で論文を検索したことになる。式が見えているため、キーワードの追加、削除など、ユーザが自由に式をカスタマイズできる点も特徴である。

式生成の対象となる文書集合を、分類コードからの検索結果に限る必要もない。任意の検索システムにおける検索結果から、任意の文書集合を選択し式を生成させ、任意の検索システムに入力するといった連携が行える。

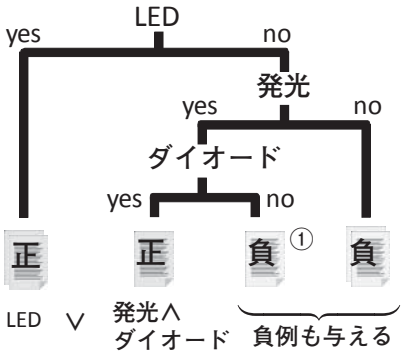
（4）その他

ある文書集合の要約を得るために、ブール式の自動生成を使うこともできる。特許マップ作成システムの多くは、システムが自動で文書をグルーピング（クラスタリング）するが、自動で集められた文書群の意味が理解しにくいことも多い。そのような場合、各グループから等価なブール式を生成すれば、そのグループの内容がとらえやすくなる。

3 提案手法

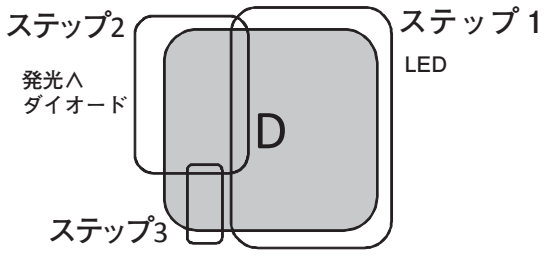
与えた文書に対してのみ真となるようなブール式を生成するには、いくつかの手法がある。まずは、カルノー図などを用いて論理式を単純化する方法がある。これは、精度が良い反面、与える文書数が多くなると速度が遅くなる。また、負例（ヒットすべきでない文書）も与えないとノイズが少ない論理式を生成することが難しい。

近年、決定木を用いて正例（与えた文書）と負例（ヒットすべきでない文書）を弁別する方法が提案された [1]。精度は犠牲になるものの非常に高速にブール式が生成できる。決定木を用いた手法の概念図を以下に示す。詳細は [1] を参照されたい。



決定木による手法の欠点は、論理式の簡略化と同じく負例を必要とする点である。ヒットすべきでない文書は与えた文書以外の文書であるから、負例の数は膨大になる。これらの負例をあまねく与えることは無理なので通常はサンプリングを行う。しかし、適度なサンプリングを行うのは難しい。特に、負例が足りないと生成した論理式が漠然としたものになってしまう。例えば、上記の決定木で①の負例を与えないと、「ダイオード」がなくても正例と負例が弁別できてしまうため、「発光∧ダイオード」ではなく「発光」のみしか生成できなくなってしまう。

そこで、正例と負例の弁別に基づく方法ではなく、正例のみからの被覆アルゴリズムに基づく方法を提案した[2]。提案手法の概念図を以下に示す。詳細は [2] を参照されたい。



ここでは、与えた文書集合が D となる。まず、この D を漏れやノイズなくカバーできる論理積を探索する。探索の目的関数は、文書検索等の評価で良く用いられている F 値で、これは再現率と精度の調和平均である。つまり、漏れが少なく（再現率が高く）ノイズも少ない（精度も高い）論理積を探索することになる。具体的には、山登り法を用いて、論理積に一個ずつキーワードを追加しながら F 値が最大となる論理積を探索する。上記の例では「LED」がまず見つかる（ステップ 1）。

一個の論理積のみで、D を漏れやノイズなくカバーすることは難しいケースも多い。提案手法では、次に、直前に生成した論理積でカバーできなかった文書集合に対して、F 値が最大となるような論理積を新たに探索する（ステップ 2）。例では「発光∧ダイオード」が見つかる。このようにして、D が空になるまで、論理積の探索を続けていく。実際は、過適合と呼ばれる現象を防ぐために、D の大きさがある一定数以下になったら探索を打ち切る。

以上、提案手法では負例を明に使っていない点に特徴がある。前述したように、負例を過不足なく与えることは難しい。ただし、論理積の探索において、負例に替わる大域的な情報を使っている。論理積の探索では、F 値を計算しているが、ここで、精度（ノイズ）の算出時に各キーワードにヒットする文書数が必要となる。精度とは、全体のヒット文書集合に占める D 中のヒット文書の割合である。

4 評価

提案手法を評価するために二種類の実験を行った。まずは、検索式の復元実験である。最初に、あるブール式から公開公報 15 年分を全文検索した。次に、検索結果から、公開日が若い順に 100 件の文書を抽出し、それらからブール式を逆生成させた。最後に、入力したブール式と逆生成させたブール式とを比較した。ブール式の生成法としては、3 節で紹介した、決定木による手法 [1] と提案手法 [2] とを比較した。結果を下表に示す。なお、決定木では、負例をランダムに 100 件選んだ。

	提案手法	決定木
一致度	0.8687	0.5276

ここで、一致度とは、0 から 1 の間の尺度で、二つのブール式が完全に一致していれば 1 に、全く異なれば 0 になる。表からもわかるように、提案手法は、決定木に比べ、30 ポイント以上一致度が高いことがわか

る。一例を挙げると、「(放熱∨(熱∧伝導)∨(伝∧熱))∧シート」から、提案手法は、「(放熱∧シート)∨(伝導∧シート)∨(伝∧熱∧シート)」を、決定木は「シート」のみを生成した。提案手法でも、入力の「熱∧伝導∧シート」に対し「伝導∧シート」までしか復元できていないが、これは、入力のブール式でヒットする 2210 件中で 100 件のみしか復元用に使っていないことが原因である。

次に、概念検索の結果からブール式を生成させ、生成させた式を評価した。具体的には、NTCIR-4 特許検索タスクの検索入力(請求項)34 件から 5 年分の公開公報を対象に概念検索を行い、検索結果の上位 100 件からブール式を生成させた。2 節の事例でも紹介したように、この 100 件のみを漏れなく検索する等価な検索式を生成することは難しい。そこで、評価では、生成したブール式が与えた 100 件の文書をどれだけ漏れなく(再現率が高く)かつノイズなく(精度が高く)検索できたかを評価した。以下に、提案手法と決定木との比較結果を示す。決定木は負例の与え方として二つの方法を試みた。まずは、ランダムに負例を 100 件選ぶ方法である。もうひとつは、[1] でも用いられている、上位 101 ～ 200 位までの 100 件を負例として用いる方法である。

	提案手法	決定木(負例:ランダム)	決定木(負例:101 ～ 200 位)
再現率	0.9271	0.9944	0.7718
精度	0.0571	0.0063	0.0627

表を見ると、いずれの手法も再現率は高いが精度が低いことがわかる。つまり、生成したブール式により、与えた上位 100 件は検索できるが、それ以外の文書も多く検索してしまうということである。決定木は、ランダムに負例を選ぶと、再現率は良いが精度が大幅に悪くなる。つまり、無駄な文書も多く検索してしまう。101 位から 200 位までを負例として与えると、今度は再現率が悪くなってしまう。ここでも負例の与え方が難しいことがわかる。

5 おわりに

概念検索と全文検索との溝を埋めるために、任意の文書集合からブール式を逆生成する手法を提案して評価した。本手法は、上記の目的以外にも、検索結果の理解支援や、異なる検索システムをつなぐためにも用いることができる。

参考文献

[1] Y. Kim, J. Seo, and W.B. Croft, “Automatic Boolean Query Suggestion for Professional Search”, In Proc. of SIGIR’11, 2011.
[2] 岩山真,「文書集合からのブール検索式自動生成」, 言語処理学会第 18 回年次大会発表論文集, pp.1336-1339, 2012.