最大クリーク探索に基づく特許検索履歴の統合

乾 孝司[1]
筑波大学大学院 システム情報工学研究科
橋本 泰一
東京工業大学 総合プロジェクト支援センター
岩山 真
日立製作所 中央研究所
難波 英嗣
広島市立大学大学院 情報科学研究科
藤井 敦
東京工業大学大学院 情報理工学研究科
橋田 浩一
産業技術総合研究所 社会知能技術研究ラボ

1　はじめに

本稿では，特許における無効資料調査や先行技術調査，特許情報に基づく技術サーベイ等の作業負荷を軽減することを目的として，これらの作業工程で必須となる特許文書検索の支援について検討する．具体的には，検索式の入力支援（クエリ入力支援）を取り上げ，ユーザが入力した検索式（クエリ）の断片に対して，それらに関連する単語群をユーザに提示することを考える．

無効資料調査等の際に実施する特許文書検索では，Web検索とは違って，検索漏れを防ぐために論理和演算子が多用されやすい．そこで本研究では，特に，論理和演算子で結合された単語の論理和結合部のクエリ入力支援に焦点をあてる．

論理和結合部の例を以下に示す．本稿では「＋」記号を使い，論理和演算を表現する．

- 例１：ウェブ＋ウエブ＋ＷＥＢ
- 例２：インターネット＋ホームページ＋ブラウザ

検索に論理和演算を利用するユーザの意図としては，例１のように，同一概念に対する表現の多様性を吸収することを目的とする場合や，例２のように，必ずしも同一の概念を指すわけではないが，互いに意味的に強い関連をもつ概念をひとまとめに検索することを目的とする場合などが考えられる．

ユーザが上記のような論理和結合クエリを作ることを考えた場合，同一概念に対する表現の多様性を吸収することを目的とする場合では，ユーザが始めに入力した単語（以降，種単語と呼ぶ）に対して，カタカナ表記をアルファベット表記に変換する，小さい「ェ」を大きい「エ」にするといった捨て仮名を変換する，長音「ー」を追加・削除するなど，幾つかの変換規則に基づいて種単語を展開することで，比較的簡単に論理和結合クエリを作ることができる．

一方で，例２にように，関連のある単語間を論理和結合するには，単語間の意味的な関連性に関する知識が必要となるが，ある単語に関連する単語（関連語）は多数存在する．また，文脈によって意味的な関連の観点が変化し，それに伴い関連語も変化する．そのため，意味的な関連性に基づいて種単語を展開することは容易な作業ではなく，某かの支援が求められている．

このような背景から，本稿では，特許文書検索においてユーザが多用する論理和結合クエリの入力を支援するために，関連語辞書を自動構築する手法を提案する．この関連語辞書は，次のような利用状況を想定している．すなわち，ユーザが論理和結合クエリを入力する際に，ユーザが入力した種単語をトリガーとして関連語辞書の辞書引きが実行され，その結果，種単語に対する関連語をユーザに提示する．

辞書構築の元となる資源として，既存の類語辞典等が挙げられる．しかし，本研究では，過去に実際に使用・蓄積された検索履歴データとして，検索クエリログに注目し，クエリログを利用して検索ユーザの経験知を直接的に辞書に反映させることを試みる．ただし，検索履歴データのみに基づいた辞書構築はあくまで初段階であり，この辞書構築の技術がある程度確立された時点で，次の段階として，類語辞典等の既に整備された言語知識との併合を目指す．

我々が構築を目指す関連語辞書の辞書項目は，見出し語と単語クラスタ群の対から形成される．単語クラスタ群の各単語クラスタは，見出し語に関連する単語のクラスタであり，見出し語との意味的関連の繋がり方によって，それぞれ別なクラスタを形成する．例として，見出し語「音声」に対する単語クラスタ群の例（一部）を以下に示す．

- 「音声」
  - クラスタ１：画像, 指紋, 虹彩, 生体, 顔, 網膜, 静脈, ...
  - クラスタ２：音楽, サウンド, 楽曲, メロディ, ...

以下本稿では，２節で，特許文書検索におけるクエリ入力支援および関連語辞書の構築に関する先行研究

[1]連絡先：inui@cs.tsukuba.ac.jp

について述べると共に，先行研究の問題点について述べる．その後，3 節で，提案する辞書構築手法について述べ，4 節で，仮想的なクエリ入力支援状況の元での利用例を述べる．

実務作業によって作成される検索式には，論理和結合部の他に，当然のことながら，論理積結合部や特許の技術範囲を区分するためのコード等が含まれている．以降本稿では，検索式から論理和結合部を抽出したデータを前提として[2]説明をおこなう．また，説明の便宜上，前後文脈から誤解がない場合，抽出された論理和結合部のことを単に検索式と呼ぶ．

2 先行研究

特許文書検索は，NTCIR のテーマ課題としても扱われており，これまでに，まとまった研究成果がある（例えば [1]）．しかし，その主な成果は，ユーザとのインタラクションが発生するようなクエリ入力支援に関するものではなく，検索アルゴリズムに関する技術に集中している．

特許文書検索に対するクエリ入力支援の基本的な手法として考えられるのは，ユーザが入力したクエリ断片と類似した特許検索履歴内の検索式をユーザにそのまま提示することである．この方法は事前準備をほぼ必要としないため簡便であるが，特許履歴の増加に伴って似たような検索式が大量に提示される．

上記の問題に対して，事前に検索履歴内の類似した検索式を統合処理し，関連語辞書や類似語辞書を構築するアプローチがある．例えば，宇野 [3] の手法は，本稿と同様，特許検索履歴の中の論理和結合部のデータから，論理和結合部の入力支援の際に参照する辞書を構築する．辞書の形式は，1 節で例示した我々の辞書と同様，項目ごとに見出し語と単語クラスタ群の対を持つ．

宇野 [3] の手法では，共通要素を持つ論理和結合部の統合処理を繰り返すことで関連語の集合を生成し，辞書の要素となる単語クラスタを作る．入力パラメータとして，統合処理の際に，どの程度の要素が共通している必要があるかを指定する．このパラメータは統合元ごとに細かな設定が可能である（設定を必要とする）．また，統合の度に，関連語集合の要素となる各単語の頻度情報（単語が統合された回数）を格納しておき，低頻度の単語を削除する等の工夫によって，関連性の低い語の混入をある程度防いでいる．

この手法は，計算コストが小さいため，大量の検索履歴に対しても比較的高速に動作する．しかし，パラメータ指定が複雑であり，また，統合処理の際に，共通要素以外の部分についてはなんら条件が課されていないため，統合が進むにつれて，関連性がそれ程高くない要素同士が同じ関連語集合に統合されてしまう危険性を有する．本稿では，この問題点に対する，ひとつの解法を提案する．

3 提案手法

3.1 基本的なアイデア

検索履歴データすなわち検索式集合から関連語辞書を構築する手法について述べる．まず，例を使って，提案手法の基本的なアイデアを説明する．いま仮に，検索履歴に含まれており，かつ共通要素をもつ以下の 2 つの検索式を統合することを考える．無条件に統合した場合，関連語集合（単語クラスタ）は $\{A, B, C, D\}$ となる．

- 検索式 1 ：$(A + B + C)$
- 検索式 2 ：$(A + C + D)$

ひとつの検索式の中で共起している単語間には，何等かの意味的関連性があると考えられる．上記の例では，検索式 1 において，A と B，B と C，および A と C が，また検索式 2 において，A と C，A と D，および C と D の間に共起関係が確認でき，各単語間には何等かの意味的な関連があると考えられる．しかし一方で，B と D は，どちらかの検索式で単独で出現しているだけであり，この単語間の関連性は他に比べて弱いことが予想される．それにもかかわらず，統合後は同じクラスタに含まれる．

本研究では，上記のような状況の場合，無条件に B と D も統合対象とするのではなく，統合条件を加えることで，より安全な統合を実現する．具体的には，検索履歴を参照し，$(A + B + D)$ や $(B + D)$ など，B と D の間の関連性を保証するような，証拠となる検索式の存在を確認する．そして，証拠の存在が確認された場合のみ，統合を許可するようにする．

ただし，このような統合条件をチェックするには，任意の検索式の統合処理の度に，統合対象以外の検索式の情報を参照する必要がある．そこで，効率良く，この処理を実現するために，グラフ理論における極大クリーク列挙問題として解くことを考える．

3.2 極大クリーク列挙に基づく検索式統合

ここで，単語をノードとし，意味的な関連性がある単語間にリンクを張ることで作られる単語グラフを考えよう．この単語グラフ上においては，リンクが密になっている部分は，相互に意味的関連性が強いことが期待される．この部分は，我々の統合処理に照らして言えば，統合が可能な部分である言える．そこで，統合対象となる検索式集合を一度，単語グラフに変換し，単語グラフ上でリンクが密になっている部分を統合することで，安全な統合を実現する．

本研究では，上記の説明の中で，「リンクが密になっ

[2]論理和結合部は，Perl や Ruby 言語を用いて正規表現規則を記述することによって，比較的簡単に抽出可能である．

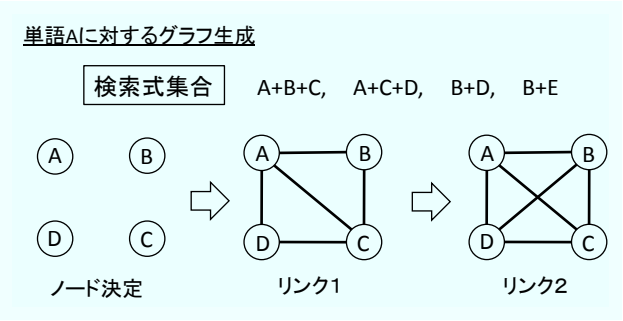

図 1: 単語グラフ生成の過程

ている部分」を「クリーク」と読み替え，検索式の統合処理問題を，単語グラフからのクリーク列挙問題として解く[3]．手続きの詳細は以下で述べるが，この処理で最終的に得られるひとつのクリークがひとつの単語クラスタに対応することになる．ひとつの単語グラフに対して，全てのクリークの列挙・抽出が並行して進行するため，統合処理の観点からみれば，以下の手続きでは，複数回の統合処理を同時並行的に処理し，単語クラスタを得ていることになる．

関連語辞書の見出し語 w ごとに単語グラフを構成し，見出し語ごとに以下のステップを実行する．

- ステップ１：単語グラフ生成
- ステップ２：単語グラフからのクリーク列挙

3.3 ステップ１：単語グラフ生成

対象単語 w に対する単語グラフ G_w を以下のように生成する．また，生成過程を図 1 に図示する．

まず，入力となる検索式集合から，単語 w を含む検索式を全て抽出する．これらの検索式に含まれる単語を G_w のノードとする．図では，対象単語を A とし，A を含む検索式 $(A+B+C)$ と $(A+C+D)$ から，４つのノードが選択されている（図左）．E は A と共起していないため選択されない．

次に，このノード間に以下の２つの手続きによってリンクを張る．なお，説明の便宜上，手続きを分割しているが，分割に本質的な意味はない．

- リンク１：上記で抽出された各検索式の中で共起している単語間にリンクを張る（図中央）．
- リンク２：抽出されない各検索式の中で共起している単語のうち，単語グラフのノードとして採用されている単語間にリンクを張る（図右）．

[3]本稿では，あるグラフ $G=(V,E)$ において，$c \subseteq V$ である c が生成するグラフが G の完全部分グラフとなっている時，c をクリークと呼ぶ．また，$u \supset c$ なる任意の u が完全グラフを構成しないとき，c を極大クリークであると呼び，さらに G の中で最大の極大クリークを最大クリークと呼ぶ．

表 1: 関連語辞書の統計量

見出し語あたりの平均単語クラスタ数	118.8
単語クラスタあたりの平均単語数	5.2

3.4 ステップ２：単語グラフからのクリーク列挙

ステップ１で生成した G_w から，G_w が含むクリークを列挙・抽出する．ただし，元の目的は検索式の統合処理であることから，全てのクリークを列挙・抽出するのではなく，極大クリークのみを列挙・抽出，あるいは最大クリークのみを抽出する．図 1 の場合，一つの極大クリーク（= 最大クリーク）$\{A, B, C, D\}$ が抽出される．そして，このクリークが単語クラスタとして出力される．入力の検索式集合には $(A+B+C+D)$ が存在しない．このことから，極大クリーク抽出によって検索式の統合処理が実現されていることがわかる．

クエリ入力支援の際には，まず，コンパクトな提示として，最大クリークを構成する単語クラスタをユーザに提示する．ユーザが追加の提示を求めた場合は，列挙した極大クリークのうち，大きなものから順次提示することによって，必要な情報を対話的に効果的に提示する．

4 関連語辞書の利用例

提案手法を用いて実データから小規模な関連語辞書を構築した．データには，一般財団法人工業所有権協力センターが 2001 年度から 2008 年度までの間に作成した検索報告書に記載された検索式約 48 万件（478,455 件）を利用し，この検索式集合から，約 33 万件（328,902 件）の論理和結合部を抽出した．抽出後の異なり単語数は 75,657，一つの論理和結合部あたりの平均結合単語数は 3.5 である．

その後，75,657 件から無作為に 100 件の単語を選択し，これらの単語を見出し語とみなして，関連語辞書の項目情報を提案手法を用いて獲得した．なお，極大クリークの列挙は Tsukiyama[2] のアルゴリズムによっておこなった[4]．また，極大クリークの中でノード数が最大のものを最大クリークとして抽出した．表 1 に獲得できた関連語辞書の統計量を示す．平均単語クラスタ数が 100 を上回っている．この値は単語の語義数に対応する値であると考えられ，市販の国語辞典等における語義の定義と比較して，粒度の細かな単語クラスタが形成されていることがわかる．

次に，仮想的なクエリ入力支援の状況を想定し，関連語辞書を利用した．まず，データとして，一般財団法人工業所有権協力センターが 2009 年度に作成した検索報告書に記載された検索式（論理和結合部）を準

[4]実装にあたり，文献 [4] の解説も参考にした．

表 2: 関連語辞書に基づくクエリ入力支援の結果例

	種単語	（上）元の検索式の情報 （下）関連語辞書によって提示された単語
正答例	個人情報 + 利用者情報	ユーザ情報　会員情報　属性情報
関連語辞書		ユーザ情報◎ ユーザデータ◯ 利用者データ◯ 顧客情報◯ 操作者情報×
正答例	画面 + ウインドウ	ウィンドウ　スクリーン　ディスプレイ
関連語辞書		ウィンドウ◎ スクリーン◎ ウィンド◯ パネル◯ 画像△ 表示△ 領域×
正答例：	ＨＴＭＬ＋ＷＥＢ	Ｗｅｂ　ウェブ　ブラウザ
関連語辞書：		Ｗｅｂ◎ ウェブ◎ ブラウザ◎ Web◯ web◯ インタネット◯ インターネット◯ ウェッブ◯ ウエッブ◯ ウエブ◯ WWW◯ ｗｅｂ◯ ハイパーテキスト◯ ホームページ◯ マークアップ◯ コンテンツ△ サイト△ リンク△ ページ× サーバ× ハイパー× 閲覧× ＨＴＴＰ× ｈｔｔｐ× ＵＲＬ×
正答例	QR コード ＋２次元コード	二次元コード　図形コード　ドットコード
関連語辞書		（なし）

備し，各検索式に含まれる単語のうち２単語を，仮にユーザが種単語として入力した単語とみなす．この状況において，検索式中の種単語以外の残りの単語を，関連語辞書を参照することによって提示できるかどうかを検証した．種単語を１単語ではなく２単語としたのは，１単語だけでは，ユーザの検索意図についての情報が乏しく，提示すべき関連語の選択が困難なためである．具体的な操作として，２つの各種単語をそれぞれ関連語辞書で辞書引きし，その後，得られた単語クラスタの中で，２つの種単語を共有している単語クラスタの構成要素を併合し，種単語以外を関連語として提示することとした．

結果例を表２に示す．種単語を２単語としたことで，ユーザの検索意図をある程度捉えることができると考えられるが，一般に，与えられた種単語に対して検索式が一意に確定するわけではないため，定量的に提示単語の良さを計測・評価することは今後の課題とする．表中の「正答例」の列が元の検索式の情報であり，「関連語辞書」の列が関連語辞書を参照することによって提示された単語群である．提示単語の右には，元の検索式にも含まれていたものに「◎」，元の検索式には含まれていないが，含まれている単語と同義であるとみなせるものに「◯」，元の検索式には含まれておらず，提示すべき関連語であるか判断に悩むものに「△」，元の検索式には含まれておらず，提示すべき関連語ではないとみなせるものに「×」をそれぞれ添えた．全般的に提示語数が多くなる傾向が観察されたが，「◎」や「◯」の単語もある程度提示することに成功していることが確認できる．

5　おわりに

本稿では，特許文書検索におけるクエリ入力支援，特に，特許文書検索においてユーザが多用する論理和結合クエリの入力を支援することを目的として，関連語辞書を自動構築する手法を提案した．辞書構築の資源として，検索履歴（検索クエリログ）に注目し，クエリログを利用して検索ユーザの経験知を直接的に辞書に反映させることを試みた．

今後は，関連語辞書自体の評価方法を検討すると共に，特許検索の実務上の要求を分析し，検索作業の負荷削減に役立つよう，関連語辞書の質の改善を進める予定である．検索履歴の利用に関しては，技術進展の速い分野では時間経過に従って使用単語も大きく変化すると考えられるため，利用履歴データの時間情報と辞書の質の関係性等について調査を進めたい．

謝辞

本研究を実施するにあたり，一般財団法人工業所有権協力センターから，特許文書検索の検索履歴データを提供して頂きました．深く感謝します．

参考文献

[1] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of the patent retrieval task at the NTCIR-6 workshop. In *Proceedings of the NTCIR-6 workshop meeting*, pp. 359–365, 2007.

[2] Shuji Tsukiyama, Mikio Ide, Hiromu Ariyoshi, and Isao Shirakawa. A new algorithm for generating all the maximal independent sets. *SIAM Journal on Computing*, Vol. 6, No. 3, pp. 505–517, 1977.

[3] 宇野喜博. 類義語統合システム. 2008. 特開 2008-152454.

[4] 宇野毅明. 大規模グラフに対する高速クリーク列挙アルゴリズム. 電子情報通信学会コンピュテーション研究会, pp. 55–62, 2003.