社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－DBS－133　(11)
2004－ＦＩ－ 75　(11)
2004／5／14

検索質問文書の主題分析に基づく類似文書検索

高木　徹†　　藤井　敦‡　　石川　徹也‡

†NTT データ 技術開発本部　〒104-0033 東京都中央区新川 1-21-2
‡筑波大学大学院 図書館情報メディア研究科　〒305-8550 茨城県つくば市春日 1-2
e-mail ：takakit@nttdata.co.jp

類似文書検索において、検索質問文書内に記述されている複数の主題要素を抽出し、主題要素ごとの検索結果と、記述特徴を考慮した主題重要度を用いた高精度な検索手法を提案する。主題要素別に、検索要求の生成、検索と主題要素重要度の付与を行い、主題要素重要度を加味した統合により最終検索結果を出力するものである。特に、本手法を特許の請求項を入力とする無効資料調査特許検索に適用する。従来の語の出現頻度の情報と、特許文書での請求項の前提部分や本質部分といった記述形式や構造情報を用いて、各構成要素重要度の算出を行う。5 年分の特許文書データを用いた評価実験により、提案手法が従来手法より高精度な検索が可能であることを示す。

Associative Document Retrieval by Query Theme Analysis

Toru Takaki†　　Atsushi Fujii‡　　Tetsuya Ishikawa‡

† Research and Development Headquarters, NTT DATA Corporation
1-21-2 Shinkawa, Chuo-ku, Tokyo, 104-0033, Japan
‡ Graduate School of Library, Information and Media Studies, University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
e-mail ：takakit@nttdata.co.jp

In this paper, we propose an associative document retrieval method that uses individual composition elements in documents by a query theme analysis. For each composition element, a query is produced and similar documents are retrieved with the relevance score. The relevance scores weighted by the importance of the corresponding composition element are integrated to determine the final relevant documents. We also propose a method for determining the importance of each composition element. We applied the proposed method to invalidity patent search. This method uses the conventional word frequency value and document-structure features, such as the preamble and essential portions in patent claim description. We evaluated our proposed method experimentally using five years worth of Japanese patent documents, and the results showed that our method was effective.

1. はじめに

知的財産の重要性が高まり、特許に関する審査や異議申し立ての迅速化が望まれている。発明が記述された文書は、その新規性や進歩性が審査機関（日本では特許庁）によって審査されたのちに、特許として成立する。その審査過程では、類似する特許や文献の検索や調査を行う先行技術調査が行われており、過去の公知の膨大な文書の中から調査を行うことが求められている。この先行技術調査において、類似資料が検索された場合には、特許として成立しない場合が多い。また、他社の製品やサービスに関連するタイムリーな特許権利調査も企業活動において重要である。

ここで、特許文書は、特有の文書構造を持ち、特許請求の範囲（請求項）、発明の属する技術分野、発明が解決しようとする課題、実施例等の項目分け記述により構成されている。特許審査官は通常、審査対象の特許文書の請求項に記述された内容について、先行技術調査を実施した上で、新規性や進歩性を判断し、特許として成立させるかを判断する。特許の審査基準では、発明が先行技術と同一か否かは、「請求項に係る

発明」であると規定されている。また、請求項は、発明の要件である動作特徴や構成特徴といった構成要素が記述されている非常に重要な部分である。本研究では、特許文書の先行技術調査において、当該特許を無効化できる過去の特許を検索する無効特許調査を対象とする。また、検索システムへの入力として、調査対象とする請求項のテキスト全文を検索質問文書として入力とすることを前提とする。これは、利用者が検索語選択の過程を経ずに検索条件の指定ができるというメリットがある。

特許文書を対象とした検索については、SIGIR2000 や ACL2003 で特許検索に関するワークショップが開催されており、情報検索の研究者の間でも重要性が認識されてきている[5,6]。また、NTCIR でも特許検索タスクとして、先行出願特許の調査を目的とする特許検索をタスクとした評価型ワークショップが開催されている。2001 年から 2002 年にかけて開催された NTCIR-3 では、新聞記事に掲載された技術や商品に関連する特許を検索する異種データ横断検索がタスクとして設定され、最初の大規模特許検索テストコレクションが構築された[4]。また、NTCIR-4 (2003 年から 2004 年に開催)では、本研究と同様に既存の特許に含まれる請求項を検索要求として，そこで請求されている権利を無効化できる特許文書を検索する無効特許調査がタスクとして設定されている[14,15]。

技術文書や特許等の知的財産文書では、発明者や研究者は新しい発明や発見を主題として記述する。この主題はこれらの文書において重要かつ不可欠な記述になっている。これらの文書は多くの場合、複数の主題が記述されており、無効特許調査の場合、検索者は検索対象とする文書の中からすべてあるいは代表的な主題を含む文書を検索したいという要求がある。さらに、検索された文書の中に、複数ある主題のうちどの主題に関する記述があるか否かを検索システムが提示できれば検索者にとって非常に効率的である。従来の検索システムでは、検索要求内の複数の主題の区別は不可能であった。

また、複数の文書内の主題は、その主題はすべて同じ重要性であるのではなく、主題によってその重要性は異なっている。検索モデルにおける検索語に対する重み付けは、その主題別に必要となってくる。特許文書の請求項での主題は、たとえば、化学分野特許では物質や化合物、機械分野では部品・装置・手段等であり、請求項における構成要素となっている。さらに例を示すと、「～するＡ手段と、～するＢ手段と、～するＣ手段とを有することを特徴とするＤ装置」という請求項表現では、「～するＡ手段」、「～するＢ手段」等が構成要素となる。

また、特許請求項の一般的な記述形式として、ジェプソン形式がある[12]。このジェプソン形式は、従来技術や構成を説明する前提部分と、特にその請求項での特徴を説明する本質部分の二つのタイプの記述形式により行われる。本質部分に属する請求項の構成要素は、前提部分に属する構成要素よりも重要であると考えられる。先行技術調査を目的とする特許検索システムでは、本質的な新規部分に着目した的確な検索が必要となる。特許検索では検索対象となる文書数が多いため、ある検索語を含む特許文書は非常に多いが、実際に本質的な内容が一致する文書は少ない。そのため、特許検索を行う場合に、入力対象の特許文書の請求項部分から、本質部分を特定して検索条件を構築することは重要である。

本研究では、特許請求項を検索質問文書とし、その主題分析を行い、主題要素として構成要素を抽出して類似特許文書検索を行う手法を提案する。文書内の構成要素ごとに検索要求の構築と検索処理を行い、各検索対象に対して検索スコア付与を行う。次に、構成要素ごとに付与された文書の検索スコアを、各構成要素の重要度を加味して統合することにより、最終検索結果を決定する。

また、各構成要素の重要度の決定方法について提案する。この方法は、従来の単語出現頻度の値に加えて、請求項記述における前提部分、本質部分という請求項特徴を用いたものである。また、５年分の特許明細書を検索対象とする評価実験により、提案手法の効果を測定する。

本稿で提案する主題分析に基づく類似文書検索手法を適用した無効特許検索システムの構成を２章で示し、３章で、請求項内の構成要素を用いた検索手法を提案する。また、４章で従来手法である構成要素を考慮しない手法と提案手法の比較を特許文書テストコレクションを用いて行う。５章で従来手法を説明し、６章でまとめる。

2. 無効特許検索システム

本章では、無効特許検索システムの処理概要を説明する。処理手順を図１に示す。無効資料調査を行う特許の請求項をシステムへの入力とし、入力に対する類似特許文書のランキングリストがシステムの出力となる。システムは入力請求項のテキスト全文から検索条件を構築し、検索処理を行う。各処理の概要を次に説明する。

Step 1 － 構成要素抽出

構成要素は入力請求項のテキスト全文から抽出する。請求項は通常典型的な記述形式により記述されているため、発明の構成要素は、請求項の記述特徴を用いたパターンのマッチング処理により自動的に抽出することが可能である [12]。構成要素抽出対象である請求項テキストに対して、形態素解析、パターンマッチングによる形態素の意味種別の付与、文脈自由文法を利用した意味種別付与された形態素間関係の特定、の各処理を順次行い、構成要素抽出を行う。次に、構成要素抽出処理の具体例を図 2 に示す。まず、構成要素抽出対象のテキストの形態素解析を行う。そして、各形態

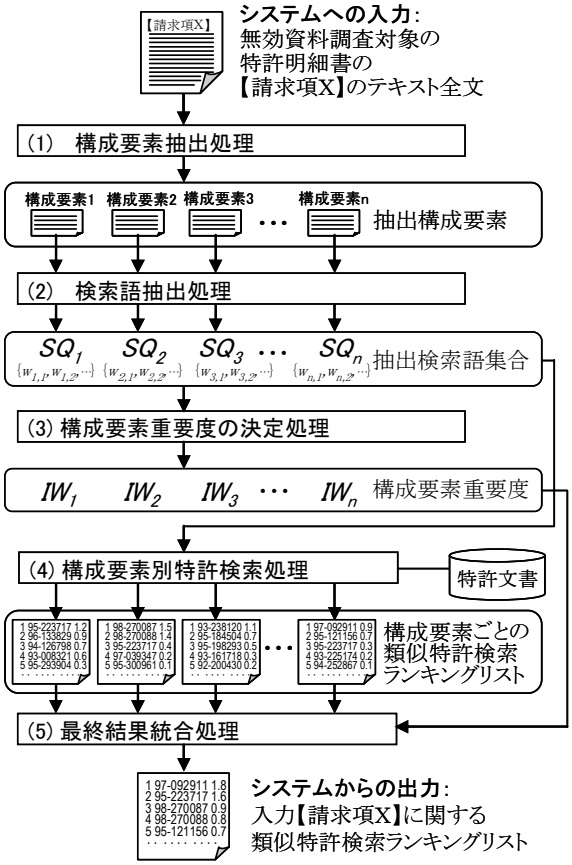

図 1　無効特許検索システムの処理手順

素の品詞、表記や漢字・平仮名・片仮名といった字種による情報の出現情報に関する正規表現に準じたパターンを用いて、形態素の意味種別情報の付与を行う。意味種別の構成要素名称を抽出するパターンでは、あとに続く形態素が、表記「と」と表記「、」である連続する名詞を構成要素名称として抽出する。次に、パターンに適合する意味種別を付与した形態素に対して、別のパターンを用いることによって、連続する形態素を一つの構成要素として抽出する。

本研究では、構成要素抽出は 241 個の人手で作成した抽出パターンを用いている。形態素解析と構成要素抽出パターンマッチングエンジンとして、それぞれ茶筌[1]と Erie システムを用いた[1] 。構成要素記述形式である前提部分および本質部分の構成要素種別を特定する処理もパターンマッチング処理により行う。

Step 2 － 検索語抽出

構成要素ごとに主に名詞単語を検索語として抽出する。さらに、連続する抽出検索語を複合語の検索語として抽出する。特に請求項に頻出する 73 語（「具備」、

[1] http://chasen.aist-nara.ac.jp

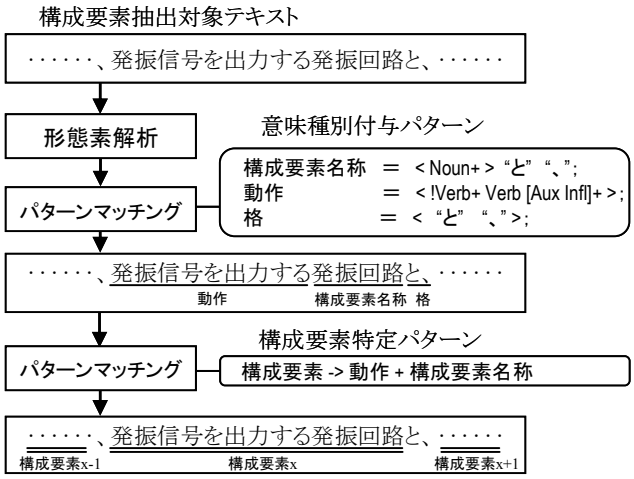

図 2　構成要素抽出処理

「請求項」、「特徴」などの語）はストップワードとして検索語から除外する。

Step 3 － 構成要素重要度の決定

抽出された各構成要素に対して相対的な構成要素重要度を付与する。この重要度は、検索結果統合処理において、構成要素ごとの検索結果を統合するときに使用する。３章で構成要素重要度の決定手法の詳細を説明する。

Step 4 － 構成要素別特許検索とランキング

構成要素ごとに入力請求項に類似する特許文書が検索対象特許文書から検索され、類似スコアが付与される。この検索ランキング処理では、従来のランキングモデルを利用する。検索結果であるランキングリストは構成要素ごとに出力される。

Step 5 － 検索結果統合

Step 4 で得られた各構成要素に基づく検索結果を統合して、類似特許文書の最終結果リストを生成する。Step 3 で算出した構成要素重要度を用いて、統合処理を行う。

3. 主題分析に基づく類似文書検索モデル

本章では、主題分析により抽出した構成要素を用いた類似特許検索モデルについて述べる。

想定する無効特許検索システムの特徴は、入力請求項の主題を構成要素として抽出し、類似する構成要素に関する内容を含む文書を検索し、さらに、各構成要素の重要度に基づいて最終検索ランキング結果を出力することである。

この類似文書検索処理において、各構成要素の重要度は異なっているため、この重要な構成要素を特定し、

これらの構成要素に大きい重要度を付与することが可能であれば、重要な主題を含む文書に対して大きな文書スコアを付与することができる。そのため、従来の一般的な検索手法に比べて提案手法はより高精度の検索が可能であり、特許調査作業等の検索効率を向上させることは可能である。

3.1 全体類似検索モデル

入力請求項から複数の構成要素を抽出し、構成要素ごとに検索要求を生成する。システムはまず構成要素別の検索要求に類似する文書を検索する。

検索要求 Q に対する類似文書 D の検索スコアを $Score(D, Q)$とする。ここで、構成要素を用いた検索モデルを式(1)のように定義する。

$$Score(D,Q) = \sum_{i=1}^{n} \left(Subscore(D, SQ_i) \times IW_i \right) \quad (1)$$

ここで、 n は構成要素の数、 SQ_i は i 番目の構成要素から生成された構成要素の検索要求 (検索語集合) 、$Subscore(D, SQ_i)$ は検索要求 SQ_iに関する文書 D の検索スコア、IW_iは i 番目の構成要素の重要度である。

3.2 文書検索モデル

文書ランキング処理で用いる検索モデルを説明する。この検索モデルは既存の一般的な検索モデルを利用する。

各構成要素の検索要求に類似する特許を検索してランキングをするために、本研究では、式(2)に示す Okapi システム[8,9] の BM25 式を用いた。

$$BM25(D,Q) = \sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf} \quad (2)$$

ここで、 Q は検索語集合 T を含む検索要求、$w^{(1)}$ は検索要求 Q 内の検索語集合 T の Robertson/Sparch Jones 重要度、k_1、 b、k_3 はパラメータ定数、 K は $k_1\left((1-b)+b\frac{dl}{avdl}\right)$、 tf は検索対象文書内の検索語出現頻度、qtf は検索要求 Q 内の検索語出現頻度、 dl と $avdl$ はそれぞれ文書長および検索対象文書集合の平均文書長である。

式(1)の $Subscore(D,SQ_i)$ は、BM25 式に基づいて式(3)のように算出される。

$$Subscore(D, SQ_i) = BM25(D, SQ_i) \quad (3)$$

BM25 関数による文書スコアの算出において、文書内に出現する検索語の重要性が高い場合と、出現頻度が大きい場合に高い文書スコアが付与される。また、構成要素内の検索語が多い場合は、検索語ごとのスコアが加算されるため、文書スコアは大きくなる傾向がある。そのため、各構成要素重要性を構成要素の文書スコアにより、直接比較することはできないという問題がある。この問題を解決する手法を 3.3 節および 3.4 節で説明する。

3.3 構成要素重要度

ここでは、式(1)の構成要素重要度(IW_i) の算出手法を説明する。

構成要素重要度は構成要素ごとに算出される値である。各構成要素に対する文書スコアは、他の構成要素とは独立に算出される。これらの重要度は各構成要素の絶対的な重要度として、正規化した値を算出するものである。

本節では、構成要素の種別に基づいた構成要素重要度の算出手法を提案する。i 番目の構成要素の重要度値 IW_iは式(4)のように算出する。

$$IW_i = 1 - CV_i \times \alpha \quad (0 \le CV_i \le 1,\ 0 \le \alpha \le 1) \quad (4)$$

ここで、 CV_i は i 番目の構成要素の重要度補正値、 α は補正値 CV_i の IW_i に対する影響度合いを変更するパラメータである。より重要であると判断される構成要素に対しては、小さな補正値 CV_iが付与される。

3.4 構成要素種別による構成要素重要度付与

構成要素種別を用いた構成要素重要度の算出手法を説明する。１章で述べたように、請求項のジェプソン記述形式は、前提部分と本質部分といった２種類の記述部分がある。特許の無効資料調査においては、前提部分の記述に比べて、本質部分の記述の方が重要であり、請求項から抽出される検索語についても、その重要度は本質部分の重要度が高いと考えることができる。前提部分に出現する語は、あまり重要でないにもかかわらず、多くの場合、本質部分でも繰り返し出現する。本質部分の記述が請求項内ではそれほど重要でない場合であっても、大きい文書スコアが付与されることになる。そのため、構成要素重要度を補正する必要がある。

前提部分を特定することは、特に日本語の特許請求項では比較的容易である。通常、前提部分の最後には、「～において」や「～であって」という表現（本稿では、前提部分終端表現と呼ぶ）が用いられており、前提部分は高い精度で自動抽出することができる。

各構成要素に対して、前提型または本質型の構成要素種別を付与する。前提型構成要素は、上述した前提部分終端表現が出現するまでの構成要素とし、それ以外の構成要素は、本質型構成要素とする。前提部分の終端表現が請求項に出現しない場合には、請求項内のすべての構成要素は本質型構成要素とする。

検索語が抽出された構成要素の種別により、構成要素補正値を決定する。これは、前提部分から抽出された検索語が、本質部分にも出現していた場合に、本質

表 1 入力請求項の構成要素の統計量

	構成要素全体			前提部構成要素			本質部構成要素		
	平均	最高	最低	平均	最高	最低	平均	最高	最低
構成要素の数	7.5	16	3	3.5	10	1	4.0	11	2
構成要素内の文字数	235.2	675	69	89.4	336	6	145.7	439	33
構成要素内の形態素数	130.6	309	36	48.6	217	3	82.0	269	18
構成要素から抽出される検索語数	111.9	256	28	26.4	80	1	85.6	224	22

部分の構成要素の文書スコアからその影響を除外するものである。本質部分の構成要素において、前提部分にも出現する検索語の数の割合を補正値を算出するために用いる。補正値は、本質部分にのみに出現する検索語が少ない場合に大きな値となり、構成要素の相対的な重要度は逆に小さくなる。

本稿では、式(5)および(6)に示す２つの補正値算出方法 *CV1* 、*CV2* を補正値算出に用いる。それぞれの算出手法を評価実験によりそれぞれの効果を比較する。補正値は、構成要素ごとに算出される。補正値 *CV1* は、構成要素における前提部分出現検索語の数の割合を適用したものである。また、*CV2* は、各検索語の *idf* を語の重要度と考え、補正値を算出したものである。各補正値 *CV1* 、*CV2* はそれぞれ 0 から 1 までの値を取り、前提部分の構成要素の場合は、1 となる。

$$CV1_i = \frac{|SQ_i \cap P|}{|SQ_i|} \tag{5}$$

$$CV2_i = \frac{\sum_{w_{i,j} \in SQ_i \cap P} idf(w_{i,j})}{\sum_{w_{i,j} \in SQ_i} idf(w_{i,j})} \tag{6}$$

$$idf(w_{i,j}) = \log\left(\frac{N}{n_{i,j}}\right) \quad (n_{i,j} > 0 \text{のとき})$$

$$idf(w_{i,j}) = \log(N) \quad (n_{i,j} = 0 \text{のとき})$$

ここで、 SQ_i は i 番目の構成要素から抽出された検索語集合、 $|SQ_i|$ は、SQ_i に含まれる検索語の数、 $w_{i,j}$ は SQ_i 内の検索語、P は前提部分構成要素に出現する検索語集合、 $|SQ_i \cap P|$ は SQ_i および前提部分構成要素に含まれる検索語の数、 N は検索対象文書の数、 $n_{i,j}$ は、語 $w_{i,j}$ を含む検索対象文書の数である。

4. 評価実験

4.1 評価方法

提案手法の効果を評価実験により検証する。評価用のデータセットとして、1993 年から 1997 年に公開された特許文書（公開特許公報）5 年分を用いた。約 170 万件の公報を含み、検索対象となるテキスト部分は約 24.8GB である。このデータセットは NTCIR-4 の特許検索タスク[14,15]で用いられているものと同じものである。情報検索のテストコレクションは通常、データセット、検索要求、正解セットからなるものであるが、NTCIR-4 特許検索タスクは進行中のプロジェクトであるため、質問要求および正解セットは完成していない。本評価では、独自に質問要求セットと、正解セットの構築を行った。質問要求セットは、過去に実際に特許庁に審査請求を行い拒絶された特許文書を 100 件収集した。また、この拒絶された特許の拒絶理由通知に記載されている拒絶の理由として引用されている先行出願特許を、検索すべき正解特許とした。正解特許を含む特許文書の最初の請求項（請求項１）をそれぞれ入力請求項（質問要求）とした。特許文書のうち、書誌情報部分を除く部分を検索対象とした。また、質問要求の各入力請求項に関する統計量を表１に示し、各質問要求に対する正解文書の数の分布を表２に示す。平均正解特許文書数は、2.09 件である。

表 2　正解文書数の分布

正解文書数	入力請求項数
1	42
2	29
3	15
4	9
5	3
6	1
7	1
平均	2.09

評価指標として、平均適合率(MAP: Mean Average Precision)を用いた。従来手法のベースラインシステムとして、構成要素抽出を行わず、入力請求項のテキスト全文から検索語を抽出する一般的な類似文書検索手法を適用した。これは、構成要素重要度を考慮せず、構成要素が一つである請求項を入力とした検索と同等である。ベースラインシステムでの検索やランキング処理は、請求項構成要素を考慮した場合と同じ検索モデル(BM25)により実行した。提案手法に関しては、式(4)で示した２つの重要度補正値を適用した。また、パラメータ α を 0 から 1 まで変化させることにより、補正値の影響度合いによる検索精度の相違を評価する。また、式(1)の BM25 関数のパラメータは、k_1=1.2、

請求項の例

サーバシステムとクライアントシステムとを有する分散オブジェクトシステムのソケットの解放装置において、前記クライアントシステムは、オブジェクトリファレンスを格納する変数の解放を検出するオブジェクトリファレンス解放検出手段と、該オブジェクトリファレンス解放検出手段により最初のオブジェクト参照時に記録したオブジェクトリファレンスとソケット番号の対応情報とを照合して不要となったソケットを検出する不要ソケット検出手段と、該不要ソケット検出手段により検出された前記不要となったソケットを切断するクライアント側不要ソケット切断手段と、該クライアント側不要ソケット切断手段により切断された前記不要となったソケットの情報を前記サーバシステムに通知する不要ソケット切断情報通知手段とを有し、前記サーバシステムは、前記不要ソケット切断情報通知手段により切断された前記不要となったソケットの情報を通知された時に、前記不要となったソケットの情報に対応するソケットを切断するサーバ側不要ソケット切断手段とを有すること特徴とする分散オブジェクトシステムのソケットの解放装置。

構成要素番号	抽出構成要素文字列	抽出検索語
1 [前提部分]	サーバシステム	サーバ
2 [前提部分]	クライアントシステム	クライアント
3 [前提部分]	分散オブジェクトシステムのソケットの解放装置	分散、解放、ソケット、オブジェクト、分散オブジェクト
4 [本質部分]	オブジェクトリファレンスを格納する変数の解放を検出するオブジェクトリファレンス解放検出手段	変数、解放、検出、格納、オブジェクト、リファレンス、オブジェクトリファレンス
5 [本質部分]	該オブジェクトリファレンス解放検出手段により最初のオブジェクト参照時に記録したオブジェクトリファレンスとソケット番号の対応情報とを照合して不要となったソケットを検出する不要ソケット検出手段	不要、最初、解放、対応、情報、オブジェクト、リファレンス、番号、照合、記録、参照、検出、ソケット、対応情報、ソケット番号、オブジェクト参照、オブジェクトリファレンス
6 [本質部分]	該不要ソケット検出手段により検出された前記不要となったソケットを切断するクライアント側不要ソケット切断手段	不要、クライアント、切断、検出、ソケット
7 [本質部分]	該クライアント側不要ソケット切断手段により切断された前記不要となったソケットの情報を前記サーバシステムに通知する不要ソケット切断情報通知手段	不要、サーバ、クライアント、通知、情報、切断、ソケット
8 [本質部分]	前記サーバシステムは、前記不要ソケット切断情報通知手段により切断された前記不要となったソケットの情報を通知された時に、前記不要となったソケットの情報に対応するソケットを切断するサーバ側不要ソケット切断手段	不要、サーバ、通知、対応、情報、切断、ソケット
9 [本質部分]	こと特徴とする	（検索語なし）
10 [本質部分]	分散オブジェクトシステムのソケットの解放装置	解放、オブジェクト、分散、ソケット、分散オブジェクト

図 3　抽出構成要素、抽出検索語の例

b=0.75、k_3=1000 とした。今回の評価セットでは、各質問要求に対する正解特許文書数が少ないため、MAP による評価は不安定である可能性がある。信憑性のある評価結果を得るためには、質問応答の研究分野で検討されたように質問要求の数を増加させることが一つの解決手法であると考えられる[13]。今回の評価では、既存の情報検索用テストコレクションの質問要求数に比べて多い 100 件の請求項を質問要求とした。

4.2 評価結果

図 3 に入力検索要求の請求項、抽出構成要素、抽出検索語の例を示す。請求項は一文で記述されており、10 個の構成要素に分割されている、最初の 3 つの構成要素は前提部分の構成要素であり、残りの 7 つは本質部分の構成要素である。抽出検索語のうち下線がある

表 3　評価結果

Parameter α	MAP		
	Baseline	CV1	CV2
0.0	0.1276	0.1306	0.1306
0.1		0.1314	0.1313
0.2		0.1319	0.1293
0.3		0.1360	0.1293
0.4		0.1367	0.1368
0.5		**0.1368**	**0.1369**
0.6		0.1334	0.1331
0.7		0.1341	0.1321
0.8		0.1284	0.1259
0.9		0.1155	0.1150
1.0		0.0884	0.0885

ものは、前提部分の構成要素にも出現している検索語である。ベースラインシステム(baseline) 及び提案手法の *CV1* および *CV2* を適用した場合の評価結果を表 3 に示す。ここで、α =0 のときは、構成要素重要度の補正を行わない場合である。ベースラインシステムの MAP 値は、0.1276 であり、構成要素の補正値の算出手法 *CV1* または *CV2* を適用した場合の MAP 値の最大値はそれぞれ 0.1368、0.1369 であった。特に *CV1* と *CV2* の大きな相違は見られなかった。また、α =0 の場合の MAP 値はベースラインシステムより若干であるが向上しており、構成要素別に検索処理を行うことの効果があることを示している。

図 4 は補正値算出方法 *CV1* (α = 0.5 のとき)を適用した場合の提案手法とベースラインシステムの 11 点平均精度を示したものである。いずれの再現率(recall)においても提案手法はベースラインシステムを上回っている。100 件の検索要求について、ベースラインシステムの結果に対して、提案手法の MAP 値が向上していたものは 50 検索要求、同等のものは 16 検索要求で、低下していたものは 34 検索要求であった。ウイルコクソンの符号順位検定により、提案手法の MAP 値の向上の統計的有意性を確認した。ウイルコクソンの符号順位検定は、しばしば情報検索分野の評価実験の検定に用いられている[2,7] 。パラメータ α が 0.1 から 0.5 の間では、有意水準 5%で提案手法の MAP 値の向上の効果が確認でき、無効資料調査を目的とする特許検索において本提案手法が有効であることを示している。

5. 関連研究

本章では、従来の検索手法と構成要素を用いた類似文書検索の相違について議論する。

ある文書に対する類似文書を検索する場合、ユーザは文書から検索語を抽出し、検索条件を作成し、システムに入力する。運用中の特許検索システムや Web 検索システムでこの方法が主に採用されている。検索要求を自然言語文で入力する他の方法では、システムは入力文から検索語と抽出して検索処理を行う。しかし、入力文字列は長い場合には、その検索要求には複数の検索話題が含まれる可能性が高くなる。多くの場合、話題は複数の語から構成されている。複数話題についての考慮をしないで検索処理がする場合、利用者の検索意図とは異なるいくつかの語があった場合、不必要な文書に高いスコアが付与されることがある。たとえば、利用者の検索要求が「高速な紙送り」と「静寂な印字」の２つが主題である印刷装置を例に考える。２つの主題がまったく区別されない場合、主題に含まれる単語の組み合わせにより、「高速な印字」や「静寂な紙送り」に関する文書も高い検索スコアで検索される可能性がある。

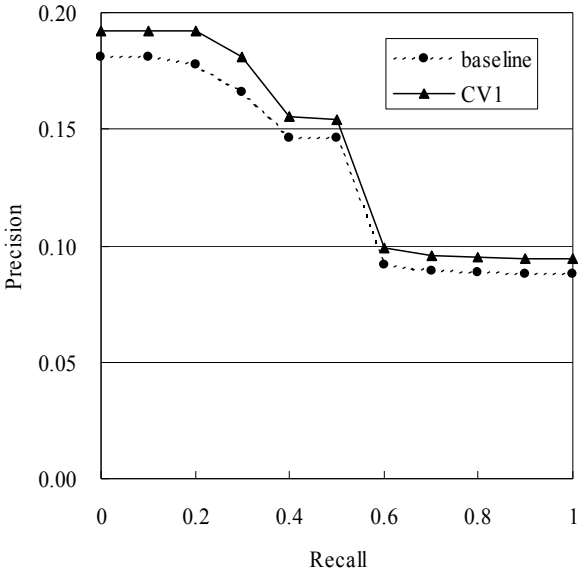

図 4　従来手法と提案手法の再現率-精度グラフ

無効資料調査を実施する際に利用可能な従来の検索システムとして、特許電子図書館(IPDL[2])等のシステムがある。これらのシステムは、検索語や国際特許分類等を論理演算子による組み合わせにより、検索を実現している。論理演算型検索システムでは、検索結果に対するランキング付与ができない問題がある。また、NRI サイバーパテントデスク[3]等の他の特許検索システムでは、検索要求として自然言語で記述された文字列を入力することが可能であり、検索結果のランキング出力が可能である。しかし、これらのシステムでは、一つの検索要求において、複数の話題に対する考慮がなく、内部的にも一つの検索要求で検索処理を実現しており、検索意図の的確な抽出が実現できているとはいえない。

また、利用者がいくつかの検索語を入力するのではなく、文書そのものを検索条件として入力とする方法として、適合性フィードバックがある[10,11]。適合性フィードバックは、我々の提案手法と、文書あるいは文書内のある一部のテキストを入力する部分では類似している。しかし、適合性フィードバックにおいても、上述した、複数話題に関する考慮がない。

従来の検索ランキング手法では、*tf-idf* 重み付けのような従来の検索語の出現頻度をベースとする手法は広く利用され、特定の語句に対する重要度の付与方法に関する検討が行われてきた。同様の考えで、重要度を付与する単位を語句から、話題や構成要素に拡張することが可能である。我々の手法は、従来の語の出現頻

[2] http://www.ipdl.jpo.go.jp

[3] http://www.patent.ne.jp

度の情報と、文書の記述形式や構造情報から、各構成要素の重要度を用いて高精度な検索を実現するものである。

6. まとめ

本稿では、類似文書検索において、検索質問文書内に記述されている複数の主題要素を抽出し、主題要素ごとの検索結果と、記述特徴を考慮した主題重要度を用いた高精度な検索手法を提案した。また、本手法を特許の請求項の各構成要素を用いた無効資料調査特許検索に適用した。この適用では、構成要素を主題要素とし、構成要素別に、検索要求の生成、検索と構成要素重要度の付与を行い、構成要素重要度を加味した統合により最終検索結果を導出した。従来から利用されている語の出現頻度の情報と、特許文書での請求項の前提部分や本質部分といった記述形式や構造情報を用いて、各構成要素重要度の算出を行った。提案手法により、重要ではない構成要素の内容を含む文書スコアを低減し、高精度な検索ランキングを実現した。５年分の特許明細書データを用いた評価実験により、提案手法が従来手法より高精度で検索可能である結果が得られた。

参考文献

[1] Y. Eriguchi and T. Kitani: NTT Data Description of the Erie System Used for MUC-6, *Proceedings of Tipster Text Program (Phase II)*, pp.469-470, 1996.

[2] D. Hull: Using statistical testing in the evaluation of retrieval experiments, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.329-338, 1993.

[3] M. Iwayama, A. Fujii, N. Kando and Y. Marukawa: An empirical study on retrieval models for different document genres: Patents and newspaper articles, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.251-258, 2003.

[4] M. Iwayama, A. Fujii, N. Kando and A. Takano: Overview of patent retrieval task at NTCIR-3, *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[5] M. Iwayama and A. Fujii, editors: *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, 2003.

[6] N. Kando: What shall we evaluate? Preliminary discussion for the NTCIR Patent IR Challenge based on brainstorming with specialized intermediaries in patent searching and patent attorneys, *Proceedings of ACM-SIGIR Workshop on Patent Retrieval*, pp.37-42, 2000.

[7] E.M. Keen: Presenting results of experimental retrieval comparison, *Information Processing & Management*, Vol. 28, No.4, pp.491-502, 1992.

[8] S.E. Robertson, S. Walker and M. Beaulieu: Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive, *Proceedings of the Seventh Text REetrieval Conference (TREC-7),* NIST Special Publication 500-242, pp.253-264, 1999.

[9] S.E. Robertson and S. Walker: Okapi/keenbow at TREC-8, *Proceedings of the Eighth Text REetrieval Conference (TREC-8),* NIST Special Publication 500-246, pp.151-161, 2000.

[10] J. J. Rocchio: Relevance feedback in information retrieval, *The SMART Retrieval System – Experiments in Automatic Document Processing,* Prentice Hall Inc., pp.313-323, 1971.

[11] G. Salton and C. Buckley: Improving retrieval performance by relevance feedback, *Journal of American Society*, Vol. 41, No.4, pp.288-297, 1990.

[12] A. Shinmori, M. Okumura, Y. Marukawa and M. Iwayama: Patent Claim Processing for Readability - Structure Analysis and Term Explanation - , *Proceedings of the ACL-2003 Workshop on Patent Corpus Processing*, pp.56-65, 2003.

[13] E. M. Voorhees and D. M. Tice: Building a question answering test collection, *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.200-207, 2000.

[14] 藤井 敦 , 岩山 真 , 神門 典子: NTCIR-4 における類似特許検索テストコレクションの構築, 情報処理学会研究報告, 2004-NL-159, pp.45-52, 2004.

[15] NTCIR-4 Patent Retrieval Task, http://www.slis.tsukuba.ac.jp/~fujii/ntcir4/cfp-en.html