検索質問の主題分析に基づく類似文書検索と特許検索への応用

高木 徹[†,††] 藤井 敦[††] 石川徹也[††]

類似文書検索において，検索質問文書に含まれる複数の主題を用いた高精度検索方式を提案する．本方式は，検索質問から抽出した主題ごとに通常の類似文書検索を行い，主題ごとに算出する主題重要度を用いて，最終的な類似文書検索結果を生成する．各主題から抽出される検索語の特定性をエントロピーを用いて算出することにより，各主題重要度を決定する．本方式を特許の請求項を検索質問とする無効特許検索に応用する．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，各主題重要度の補正を行う．NTCIR 特許文書テストコレクション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 Subtopic Analysis and its Application to Patent Search

Toru Takaki,[†,††] Atsushi Fujii[††] and Tetsuya Ishikawa[††]

In this paper, we propose an associative document retrieval method by query subtopic analysis. Our method uses the individual subtopic elements in a query document and retrieves the associative documents on a subtopic-by-subtopic basis. For each subtopic element, a subquery is produced and similar documents are retrieved with the relevance score. The relevance scores weighed by the importance of each subtopic element are integrated to determine the final relevant documents. In calculation of subtopic's importance, the specificity of a query term is evaluated using entropy, which is a deviation degree of term occurrences in each subtopic element. We applied the proposed method to an invalidity patent search in which subtopics are the composition elements in a query claim. We propose an additional calculation method of subtopic's importance using the feature of query patent claim, such as preamble and essential portions. We evaluated our method experimentally using the NTCIR patent IR test collection. The results showed that our method was effective than existing methods in retrieval accuracy.

1. はじめに

類似文書検索は，文書を検索質問とする検索方式であり，入力文書に類似した文書が出力となる．一般的な類似検索システムでは，検索質問として入力された文書から検索語を抽出して，これら検索語の出現頻度情報を用いて，類似文書の検索やランキングを行う．類似文書検索は，検索語の指定を行うことなく，検索作業効率の向上が期待できる．文書自体を検索質問として指定できない非類似文書検索システムでは，利用者は検索語を取捨選択して，検索語を指定する必要がある．

† NTT データ技術開発本部
Research and Development Headquarters, NTT DATA Corporation
†† 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

文書は，特定の主題について記述されている．技術論文や特許文書といった文書では，いくつかの技術を組み合わせて論じることが多いため，主題は複数あることが多い．技術論文では，いくつかの従来手法や提案手法が主題になる．特許文書の請求項では，発明の構成特徴や動作特徴が主題となる．

複数の主題を含む文書を検索質問として類似文書検索を行う場合，利用者には，検索質問内のすべて，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．従来の検索システムでは，検索質問内の個々の主題は区別しない．そのため，検索結果に利用者が意図しない主題を含む文書が検索される場合がある．

本研究では，複数の主題を含む検索質問文書を入力とする類似文書検索の高精度化手法を提案する．検索質問文書の主題抽出を行い，主題ごとの重要度を加味した類似度計算により類似文書検索を行う．主題重要度の算出では，各主題に出現する語の特定性に着目する．

また，本手法を特許検索に応用し，入力特許文書の特徴を主題重要度に反映させる手法を提案する．

2 章で，本論文で提案する主題分析に基づく類似文書検索手法の処理手順，主題抽出および主題重要度の算出方法の説明を行う．3 章で，本提案方式の特許文書検索への応用について説明する．4 章で，本提案手法の有効性を評価実験によって示す．5 章で，関連研究について議論する．

2. 主題分析に基づく類似文書検索

2.1 処 理 概 要

検索質問文書の主題分析に基づく類似文書検索手法について説明する．複数の主題を持つ検索質問文書において，各主題の重要度は異なっている．重要な主題を特定し，高い重要度を付与することができれば，その主題に関連する文書に対して高い文書スコアを付与することができる．その結果，従来の検索手法に比べて高精度な検索が可能となる．提案する類似文書検索の処理手順を**図 1** に示す．各処理の概要を次に示す．

Step 1 — 主題抽出

技術文書には，問題や方法が主題として記述されている．数文程度から成る短い文書では，主題は 1 つである可能性が高い．しかし，多くの文を含む文書では，主題が複数存在することがある．

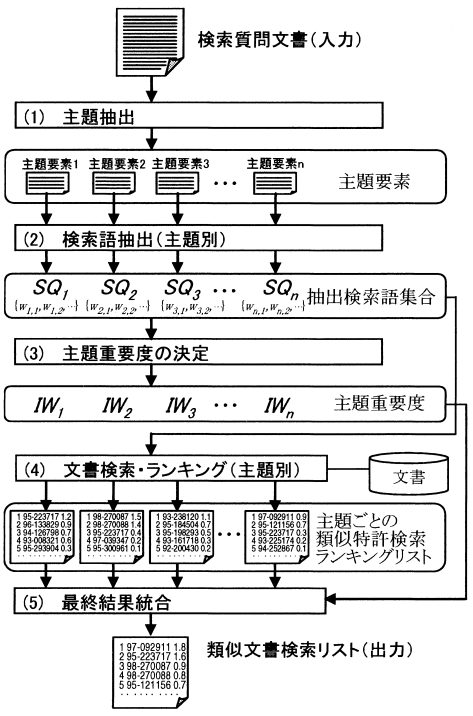

図 1 主題分析に基づく類似文書検索の処理フロー
Fig. 1 Processing flow of associative document retrieval by query subtopic analysis.

主題は，文書から抽出されるテキストとする．そこで，目的に応じて，テキストセグメンテーション手法や，文書特有の記述特徴を用いたパターンマッチングによる抽出手法を使うことができる．

Step 2 — 主題別の検索語抽出（検索質問作成）

主題ごとの類似文書検索（Step 4）を行うために，検索質問の作成を行う．検索質問は，各主題に対応するテキストから抽出した単語の集合である．

Step 3 — 主題重要度の決定

抽出された各主題に対して重要度を付与する．この重要度は，Step 5 において，主題ごとの検索結果を統合するときに使用する．2.3 節で主題重要度の決定手法の詳細について説明する．

Step 4 — 主題別文書検索とランキング

主題ごとの検索質問を用いて文書を検索し，類似スコアを付与する．ここで，従来の文書検索モデルを利用し，主題ごとに文書リストを作成する．

Step 5 — 検索結果統合

Step 4 で得られた主題ごとの文書リストを統合して，類似文書の最終検索リストを生成する．ここで，Step 3 で算出した主題重要度を用いる．

2.2 検索モデル

本研究で利用する検索モデルについて説明する．本検索モデルは，Step 4 の文書検索モデルと，Step 5 の統合モデルから構成される．2 つのモデルを Step 5, Step 4 の順で説明する．

2.2.1 統合モデル

検索質問から複数の主題を抽出し，主題ごとに検索質問を生成する．次に，主題別の検索質問に類似する文書を検索する．

検索質問 Q に対する類似文書 D の類似スコアを $Score(D, Q)$ とする．主題を考慮した検索結果統合モデルを式 (1) で定義する．

$$Score(D,Q) = \sum_{i=1}^{m} (Subscore(D, SQ_i) \times IW_i) \qquad (1)$$

ここで，m は主題要素の数，SQ_i は i 番目の主題要素から生成された主題の検索質問，$Subscore(D, SQ_i)$ は検索質問 SQ_i に関する文書 D の類似スコア，IW_i は i 番目の主題に付与された重要度である．

2.2.2 文書検索モデル

文書検索モデルには既存の一般的な検索モデルを利用することができる．ただし，本研究では，試験的に式 (2) に示す Okapi BM25 を用いた[9]．i 番目の検索質問

SQ_i に関する文書 D の類似スコア $Subscore(D, SQ_i)$ は，式 (2) で計算する．

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

ここで，$w^{(1)}$ は検索質問 SQ_i 内の検索語 T に対する Robertson/Sparck Jones 重要度，k_1，b，k_3 は定数，K は $k_1((1-b)+b\frac{dl}{avdl})$，$tf$ は検索対象文書内の検索語出現頻度，qtf は検索質問 SQ_i 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．

2.3 検索語の出現分布を用いた主題重要度算出

本研究のポイントは，重要な主題に対して大きな重要度を付与する点にある．

主題の重要度は，主題内に含まれる検索語の重要度を総和した値と考える．ここで，検索語の重要度として，逆文書頻度（IDF）が一般的に利用されている．IDF は，ある文書集合においてある検索語が含まれる文書数を用いたもので，少数の文書のみに出現する検索語に大きい値を付与する．

IDF のほかに，信号/雑音比により算出することが可能である．信号/雑音比は，検索語が文書集合での出現文書の偏り度合いを情報理論のエントロピー尺度を用いて表す．しかし，文書集合での検索語の出現頻度や出現分布を用いた従来の尺度では，同一文書内の検索語の出現分布を考慮することができない．そのため，特定性のない語に対しても高い値が与えられる場合があり，主題重要度の算出には不適切である．

本研究では，主題における検索語の出現分布を反映させて，主題に基づく相対的な重要度を算出する．多くの主題に出現する語は特定性が低く，特定の主題に出現する語は特定性が高いと考える．

検索語の重要度は，主題における出現数を，検索語の出現分布から算出したエントロピーで補正する．

次に，主題重要度の算出方法について説明する．

検索語 w_j が主題 i に対応する検索質問 SQ_i に出現する確率を p_j^i とする．p_j^i は，検索語 w_j が検索質問文書に出現する頻度と，検索質問 SQ_i における出現頻度から推定できる．主題検索質問 SQ_i 中の検索語 w_j の出現頻度を $tf_{j,i}$ とすると，式 (3) で算出する．

$$p_j^i = \frac{tf_{j,i}}{\sum_{k=1}^{m} tf_{j,k}} \quad (3)$$

各主題の検索語集合 $\{SQ_1, SQ_2, \ldots, SQ_m\}$ に検索語 w_j が出現することを表す確率変数のエントロピー n_j は式 (4) と式 (5) で計算される．

$$n_j = -\sum_{i=1}^{m} p_j^i \log_2 p_j^i \quad (4)$$

$$n_j = -\sum_{i=1}^{m} \frac{tf_{j,i}}{\sum_{k=1}^{m} tf_{j,k}} \log_2 \frac{tf_{j,i}}{\sum_{k=1}^{m} tf_{j,k}} \quad (5)$$

ここで，m は検索質問文書内の主題数である．

1 つの主題にしか出現しない検索語は $n_j = 0$ となるため，出現頻度のスムージングを式 (6) の加算法で行う．

$$n_j = -\sum_{i=1}^{m} \frac{tf_{j,i}+\delta}{\sum_{k=1}^{m}(tf_{j,k}+\delta)} \log_2 \frac{tf_{j,i}+\delta}{\sum_{k=1}^{m}(tf_{j,k}+\delta)} \quad (6)$$

ここで，δ は加算値パラメータである．検索語の全主題の出現頻度をエントロピーで補正するために，検索語 w_j の重要度 s_j を式 (7) で定義する．

$$s_j = \log_2 \sum_{i=1}^{m} tf_{j,i} - n_j \quad (7)$$

主題の重要度は，そこに現れる検索語の総和であると考え，式 (8) か式 (9) のいずれかで計算する．

$$IW1_i = \frac{1}{\log_2(1+|SQ_i|)} \sum_{w_j \in SQ_i} s_j \quad (8)$$

$$IW2_i = \frac{1}{|SQ_i|} \sum_{w_j \in SQ_i} s_j \quad (9)$$

式 (8) と式 (9) は語数による正規化の手法が異なる．これら 2 つの正規化手法の異なる重要度算出方法は，評価実験で両者の効果を測定し，比較する．

3. 特許検索への応用

2 章で提案した手法を特許検索に応用する．また，特許文書の特徴を考慮して，主題重要度の補正を行う．

3.1 無効特許検索

知的財産の重要性が高まり，特許審査の迅速化が望まれている．特許審査では，膨大な公知文書から類似する資料を検索する「先行技術調査」が行われる．類似資料が検索された場合は，原則，特許として成立しない．

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，実施例等の項目で構成されている．特許審査は，審査対象特許の請求項について先行技術調査を実施したうえで，新規性や進歩性を判断し，特許と

しての適否を決定する☆．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主題が記述されている．すなわち，請求項を検索質問とする類似文書検索は有用性が高い．

SIGIR2000 や ACL2003 で特許検索に関するワークショップが開催され，情報検索研究者の間でも重要性が認識されている[7),8)]．また，情報アクセス技術の促進を目的とした国際的な評価型ワークショップ NTCIR（国立情報学研究所主催）でも，先行出願特許の調査を目的とする特許検索タスクが行われている[2),5),6)]．NTCIR-3（2001 年から 2002 年に開催）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が行われ，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[6)]．NTCIR-4（2003 年から 2004 年に開催）では，本研究で提案する検索システムの適用分野と同じ無効特許検索タスクが行われた[2)]．

3.2 特許文書の特徴

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．

文書内において複数の主題は，すべて同じ重要性を持つわけではない．特許文書の請求項における主題は，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・手段等の構成要素である．「〜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」という請求項では，「〜する A 手段」,「〜する B 手段」等が構成要素となる．以下，特許請求項における主題を「構成要素」と呼ぶ．

NTCIR-4 特許検索タスクの検索課題の請求項の例を**図 2** に示す．⟨CLAIM⟩ タグで括られた部分が入力請求項となる．⟨COMP⟩ タグで括られた部分が構成要素である．本研究では，構成要素は独自の方法によって自動分割した．

特許請求項の記述形式として，ジェプソン形式がある[12)]．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，特に請求項での特徴を説明する「本質部分」で構成される．無効特許検索では，本質的な新規部分に着目した的確な検索が必要となるため，本質部分に属する構成要素は，前提部分に属する構成要素よりも重要である．

入力された請求項から，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て主題分析を行い，主題として構成要素を抽出する．

```
<TOPIC>
<NUM>023</NUM>
<FDATE>19970612</FDATE>
<DOC>
【特許請求の範囲】
<CLAIM>
【請求項1】
<COMP><CNUM>1</CNUM>対向する一対の基板間に挟持された液晶を駆動し、</COMP>
<COMP><CNUM>2</CNUM>その液晶により画像を表示する液晶表示装置において、</COMP>
<COMP><CNUM>3</CNUM>前記対向する一対の基板の少なくとも一方の基板のパターン空白部に、</COMP>
<COMP><CNUM>4</CNUM>穴空けもしくは切欠き加工を施したこと</COMP>
<COMP><CNUM>5</CNUM>を特徴とする液晶表示装置。</COMP>
</CLAIM>
</TOPIC>
```

図 2 特許請求項と構成要素の例

Fig. 2 Example of patent claim and extracted subtopic elements.

3.3 無効特許検索への応用

無効特許検索システムの処理について説明する．基本手順は，2.1 節で説明した類似文書検索手法と同じである．無効特許検索を行う特許の請求項を入力し，入力に対する類似特許文書のランキング付き検索リストを出力する．提案手法を無効特許検索に応用するための具体的な方法を次に説明する．

Step 1 — 構成要素抽出

主題すなわち構成要素は入力請求項から抽出する．請求項は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マッチングにより自動的に抽出することが可能である[12)]．入力請求項に対して形態素解析を行い，パターンマッチングによって形態素に意味種別を付与し，文脈自由文法によって意味種別付与された形態素間関係を特定する．

構成要素抽出処理の具体例を**図 3** に示す．まず，構成要素抽出の対象テキストを形態素解析する．各形態素の品詞，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を付与する．意味種別の構成要素名称を抽出するパターンでは，後ろに「と,」が続く連続する名詞を構成要素名称として抽出する．次に，意味種別を付与した形態素に対して，連続する形態素を 1 つの構成要素とする．

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．形態素解析には茶筌☆☆，構成要素抽出には Erie[1)] を用いた．NTCIR-4 特許検索タスクの検索課題で，人手による構成要素抽

☆ 日本の特許審査基準では，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．

☆☆ http://chasen.aist-nara.ac.jp/

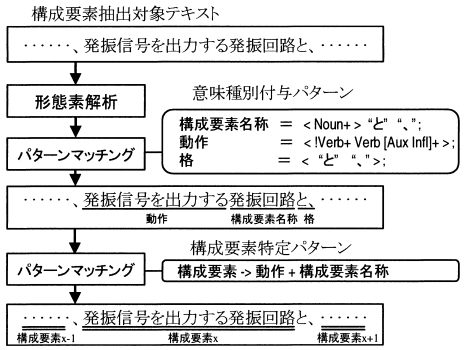

図 3 構成要素抽出処理の例
Fig. 3 Example of composition-element extraction.

出（図 2 の ⟨COMP⟩ タグで括られた部分）と，本研究の構成要素の自動抽出を比較し，構成要素の一致率を F 値で算出したところ 0.82 であり，妥当な構成要素抽出が実現できた．

前提部分および本質部分もパターンマッチングによって特定する．

Step 2 — 検索語抽出

構成要素ごとに名詞を検索語として抽出する．さらに，連続する検索語を複合語として抽出する．請求項に頻出する 73 語（「具備」,「請求項」,「特徴」等の語）はストップワードとして検索語から除外する．

Step 3 — 構成要素別特許検索とランキング

2.3 節で説明した主題重要度の算出方法を無効特許検索に適用する．ここで，構成要素に対する重要度の算出を行う（3.4 節）．

Step 4 — 構成要素重要度の決定

2.2.2 項で説明した文書検索モデルを用いて，構成要素別に特許検索処理を行う．

Step 5 — 検索結果統合

2.2.1 項で説明した主題別検索結果の統合モデルを用いて，類似文書検索の最終的な結果を生成する．

3.4 構成要素種別による主題重要度付与

3.2 節で説明したように，ジェプソン形式は，前提部分と本質部分の記述部分があり，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る．

前提部分の特定は，日本語の特許請求項では容易である．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（本論文では,「前提部分終端表現」と呼ぶ）が用いられるため，前提部分は高い精度で自動抽出することができる．抽出された各構成要素に対して，前提部分または本質部分の種別を付与する．前提部分は，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部分とする．図 2 の例では，前の 2 つの構成要素が前提部分となっている．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を本質部分とする．検索語が抽出された構成要素の種別により，構成要素補正値の補正を行う．本研究では，前提部分に含まれる構成要素の重要度を α 倍（$0 \leq \alpha \leq 1$）とする．すなわち，式 (8) と式 (9) をそれぞれ式 (10) と式 (11) で置き換える．

$$IW1_i = \frac{1}{\log_2(1+|SQ_i|)} \sum_{w_j \in SQ_i} s_j \times \alpha \quad (10)$$

$$IW2_i = \frac{1}{|SQ_i|} \sum_{w_j \in SQ_i} s_j \times \alpha \quad (11)$$

ここで，α を 1 より小さくすると，本質部分を前提部分よりも重要視することができる．また，$\alpha = 0$ のときは，前提部分をまったく考慮しない．

4. 評価実験

4.1 評価方法

提案手法の有効性を評価実験により検証する．評価用のテストコレクションとして，NTCIR-4 特許検索タスクで使用された文書セット，検索課題，適合判定を使用した☆．本テストコレクションは無効特許検索をタスクとする．文書セットは 1993 年から 1997 年に公開された特許文書（日本公開特許公報）5 年分であり，約 170 万件の文書を含み，検索対象となるテキストのファイル容量は約 24.8 GB である．また，検索課題は，上記の文書セットに正解を含む特許文書から抽出された特定の請求項である．検索課題の数は 102 件である．検索課題に対する正解は，NTCIR 参加者が提出した検索結果をプーリングした特許文書を特許専門家が適合性の判定をした文書と，専門家が独自に見つけた文書である．適合判定レベルは，単独で無効化できる特許（A），他との組合せによっては無効化できる特許（B）がある．本評価では，A および B をともに正解とした．平均正解文書数は 4.5 件である．検索課題の各入力請求項から抽出された構成要素（主題）は平均 4.8 個（最大 14 個，最小 2 個）である．

評価指標として，平均精度の平均（MAP：Mean Average Precision）を用いた☆☆．本評価では，次の

☆ http://www.slis.tsukuba.ac.jp/~fujii/ntcir4/cfp-en.html

☆☆ MAP および再現率–精度は専用のスコア計算プログラム（trec_eval）で計算した．ftp://ftp.cs.cornell.edu/pub/smart/trec_eval.v3beta.shar から入手可能．

4 手法の比較を行った．

BASE 主題抽出を行わない場合

SE 主題抽出を行うものの，主題重要度を用いない（重要度はすべて 1 とする）場合

SE+IW1 主題抽出を行い，主題重要度算出で式 (10) を用いる場合

SE+IW2 主題抽出を行い，主題重要度算出で式 (11) を用いる場合

ベースラインシステム（**BASE**）として，主題抽出を行わず，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，検索質問文書の主題を 1 つと見なして入力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，主題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（**SE+IW1** および **SE+IW2**）に関しては，それぞれ式 (10)，式 (11) で示した 2 つの主題重要度算出方法を適用した．また，パラメータ α を変化させることにより，前提部分の重要度を変化させ，特許特有の記述特徴を用いた場合の効果を測定した．さらに，主題重要度の効果を比較するために，主題重要度を適用しない場合の評価も行った（**SE**）．

本実験では，式 (2) の BM25 関数のパラメータは，$k_1 = 1.2$，$b = 0.75$，$k_3 = 1000$ とした．また，式 (6) のスムージング加算値を $\delta = 0.5$ とした．これらは一般的な値である．

4.2 評価結果

前節で説明した 4 手法（**BASE**，**SE**，**SE+IW1**，**SE+IW2**）の評価結果を**図 4** に示す．横軸は，主題重要度を適用した場合のパラメータ α である．**BASE** と **SE** は，主題重要度を適用していないため α によらず MAP は一定である．**BASE** と **SE** を比較すると，**SE** は若干 MAP が低下している．すなわち，主題抽出だけを行って結果を統合しても効果がないことが分かる．

BASE と **SE+IW1** の比較では，α が 0.1 から 0.6 のとき，すなわち，特許請求項の前提部分をあまり重視しない場合には，**BASE** を上回る MAP が得られた．$\alpha = 0.2$ のとき，MAP は最大 0.1484 であり，**BASE** に比べて向上した．

α が大きい場合，すなわち，特許請求項の前提部分と本質部分の重要度の相違を考慮しない場合は，効果がないことが分かる．

IW1 と **IW2** の比較では，重要度の正規化手法として **IW1** が有効であることが分かる．

検索課題中の特許請求項では，前提部分を持たない請求項も存在する．NTCIR-4 特許検索タスクテストコレクションでは，102 件の検索課題のうち 46 件が前提部分を持つ．前提部分を含む検索課題のみを用いた場合の評価結果を**図 5** に示す．前提部分を持つ検索課題に限定して，**BASE** と **SE+IW1** の比較を行うと，**BASE** の MAP が 0.1192 に対して，**SE+IW1** では 0.1594 と向上した☆．

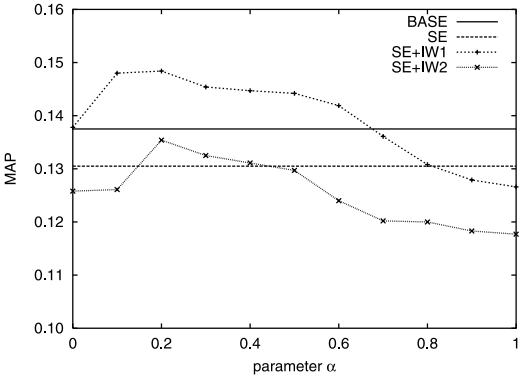

図 4 評価結果（MAP：全検索課題）
Fig. 4 Evaluation results (MAP: all topics).

図 6 は，**BASE** と **SE+IW1**（$\alpha = 0.2$ のとき）の再現率–精度グラフである．いずれの再現率（recall）においても，提案手法の精度（precision）はベースラインシステムを上回っている．本評価結果より，主題抽出と主題重要度が，無効特許検索を目的とする類似特許検索で有効であることが分かった．

算出した各構成要素重要度と，構成要素ごとに特許文書検索を行ったランキング結果の MAP の相関を測定し，構成要素重要度の妥当性を分析した．構成要素ごとに検索質問を生成し検索を行った場合に，高い MAP が得られる構成要素は重要である．すなわち，高い MAP が得られる構成要素の重要度を高くすることにより，最終検索結果の MAP を向上することができる．**SE+IW1** で，パラメータ α を 0.1 刻みに変化させたときに，構成要素重要度と構成要素別の MAP の相関係数は，平均 0.241（最小 0.208，最大 0.263）となった．前提部分を含む検索課題に限定した場合では，平均 0.249（最小 0.177，最大 0.296）であった．いずれの場合も，「やや相関あり」で，構成要素の重要度が妥当であることが分かった．

5. 関連研究

類似文書検索において，利用者は文書から検索語を

☆ 正解が 1 文書である検索課題の場合，正解の順位が 8.39 位から 6.27 位に向上したことを意味する．

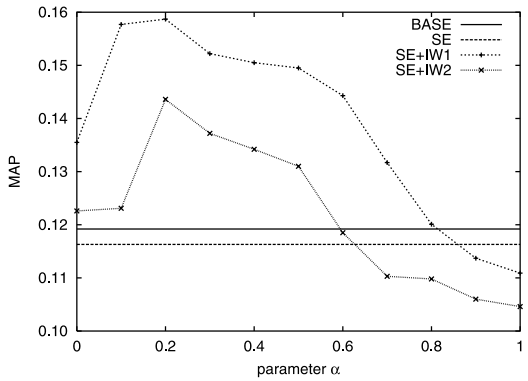

図 5 評価結果（MAP：前提部分を持つ検索課題）
Fig. 5 Evaluation results (MAP: topics having preamble portion).

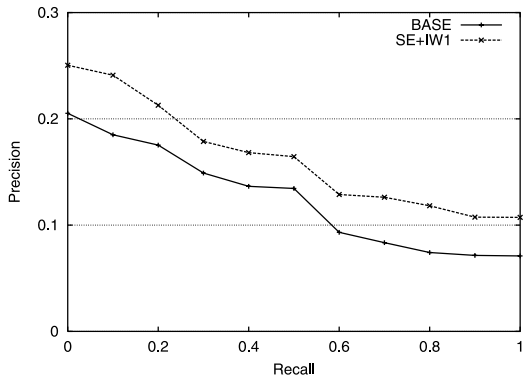

図 6 再現率–精度グラフの BASE と SE+IW1 の比較
Fig. 6 Recall-Precision curve for BASE and SE+IW1.

抽出し，検索質問を作成してシステムに入力する．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．

検索質問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を抽出して検索処理を行う．しかし，入力文字列が長い場合には，検索質問には複数の主題が含まれる可能性が高くなる．また，主題は複数の語から構成されているため，複数主題の考慮をせずに検索処理を行うと，利用者の情報要求とは無関係な語があった場合に，不要な文書に高いスコアが付与される．たとえば，利用者の検索質問として「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に関する文書を考える．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合せにより，「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．本研究の提案手法は，主題ごとに検索することが可能であり，この問題を解決することができる．

解説的文書が主題で構成されていることに着目し，検索対象文書の主題を利用した検索や，検索対象文書の局所的な情報を考慮したパッセージ検索がある[3),4),14)]．しかし，本研究は，検索質問文書の主題分析を行う点が異なる．

また，利用者が文書そのものを検索質問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，文書を入力する点で本研究と類似している．しかし，適合性フィードバックは複数の主題を区別しない．

従来の文書検索手法では，検索語の出現頻度に基づく手法が利用されている．同様の考え方で，重要度を付与する単位を語句から，主題や構成要素に拡張することが可能である．本手法は，従来の語の出現頻度の情報に加えて，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．

既存の特許検索システムとして，特許電子図書館（IPDL☆）等がある．利用者は，検索語や国際特許分類等を用いて論理式を構成し，検索を行う．しかし，論理型システムでは，検索結果の順位付けができない問題がある．NRI サイバーパテントデスク☆☆等の特許検索システムでは，自然言語による検索が可能である．しかし，検索質問における複数の主題を区別することができない．本手法は，特許検索システムに対しても有用であり，特許調査作業の効率化に寄与することができる．

6. ま と め

本論文では，検索質問文書内に記述されている複数の主題を抽出し，主題の重要度を用いた高精度な類似文書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特許検索に応用した．

この応用では，構成要素を主題とし，構成要素別に，検索質問の生成，検索と構成要素重要度を加味した統合を行い，最終検索結果を生成した．さらに，特許請求項で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，構成要素の重要度を算出し，提案手法により，重要ではない構成要素に関連する文書のスコアを低減させ，高精度な検索を実現した．評価実験により，無効特許検索において，提案手法が従来手法より高精度で検索可能であるという結果が得られた．

検索結果の提示の際，どの主題に関連する文書か否かを表示することにより，検索結果の内容理解を容易にするほかの効果もある．また，主題抽出において情

☆ http://www.ipdl.ncipi.go.jp/
☆☆ http://www.patent.ne.jp/

報抽出やテキストセグメンテーションを利用することにより，特許文書以外の検索へも応用可能である．

参 考 文 献

1) Eriguchi, Y. and Kitani, T.: NTT Data Description of the Erie System Used for MUC-6, *Proc. Tipster Text Program* (*Phase II*), pp.469–470 (1996).
2) Fujii, A., Iwayama, M. and Kando, N.: Test Collection for Patent-to-Patent Retrieval and Patent Map Generation in NTCIR-4 Workshop, *Proc. 4th International Conference on Language Resources and Evaluation*, pp.1643–1646 (2004).
3) Hearst, M.A. and Plaunt, C.: Subtopic Structuring for Full-Length Document Access, *Proc. 16th Annual International ACM SIGIR Conference*, pp.59–68 (1993).
4) Hearst, M.A.: Multi-Paragraph Segmentation of Expository Text, *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*, pp.9–16 (1994).
5) Iwayama, M., Fujii, A., Kando, N. and Marukawa, Y.: An empirical study on retrieval models for different document genres: Patents and newspaper articles, *Proc. 26th Annual International ACM SIGIR Conference*, pp.251–258 (2003).
6) Iwayama, M., Fujii, A., Kando, N. and Takano, A.: Overview of patent retrieval task at NTCIR-3, *Proc. 3rd NTCIR Workshop* (2003).
7) Iwayama, M. and Fujii. A., editors: *Proc. ACL-2003 Workshop on Patent Corpus Processing* (2003).
8) Kando, N.: What shall we evaluate? Preliminary discussion for the NTCIR Patent IR Challenge based on brainstorming with specialized intermediaries in patent searching and patent attorneys, *Proc. ACM-SIGIR Workshop on Patent Retrieval*, pp.37–42 (2000).
9) Robertson, S.E. and Walker, S.: Okapi/keenbow at TREC-8, *Proc. 8th Text REetrieval Conference* (*TREC-8*), pp.151–161 (2000).
10) Rocchio, J.J.: Relevance feedback in information retrieval, *The SMART Retrieval System — Experiments in Automatic Document Processing*, pp.313–323, Prentice Hall Inc. (1971).
11) Salton, G. and Buckley, C.: Improving retrieval performance by relevance feedback, *Journal of American Society*, Vol.41, No.4, pp.288–297 (1990).
12) Shinmori, A., Okumura, M., Marukawa, Y. and Iwayama, M.: Patent Claim Processing for Readability — Structure Analysis and Term Explanation, *Proc. ACL-2003 Workshop on Patent Corpus Processing*, pp.56–65 (2003).
13) Voorhees, E.M. and Tice, D.M.: Building a question answering test collection, *Proc. 23rd Annual International ACM SIGIR Conference*, pp.200–207 (2000).
14) Wilkinson, R.: Effective Retrieval of Structured Documents, *Proc. 17th Annual International ACM SIGIR Conference*, pp.311–317 (1994).

(平成 16 年 10 月 18 日受付)
(平成 17 年 2 月 1 日採録)

高木 徹（正会員）

1990 年筑波大学第三学群情報学類卒業．1992 年同大学大学院修士課程理工学研究科修了．同年 NTT データ通信（株）（現（株）NTT データ）入社．情報検索，自然言語処理の研究開発に従事．現在，筑波大学大学院図書館情報メディア研究科博士後期課程在学中．ACM，日本データベース学会各会員．

藤井 敦（正会員）

1993 年東京工業大学工学部情報工学科卒業．1998 年同大学大学院博士課程修了．図書館情報大学助手を経て，現在，筑波大学大学院図書館情報メディア研究科助教授．博士（工学）．自然言語処理，情報検索，音声言語処理，Web マイニングの研究に従事．電子情報通信学会，人工知能学会，言語処理学会，Association for Computational Linguistics 各会員．

石川 徹也（正会員）

1977 年慶應義塾大学大学院修士課程（図書館情報学専攻）修了．富士フイルム（株）足柄研究所，図書館短期大学，図書館情報大学を経て，現在，筑波大学大学院図書館情報メディア研究科教授．工学博士．情報管理システムの高度化に関する研究に従事．人工知能学会，言語処理学会，ACM 等各会員．