文書構造と共起表現を用いた文書ランキング手法

野本 昌子 野口 直彦
松下電器産業（株） マルチメディアシステム研究所

1 はじめに

ベクトル空間モデル，確率モデルなどの非完全一致モデルに基づく情報検索手法[1]では，何らかの類似尺度によって文書のランキングを行うが，その類似計算には単語の頻度情報や分布情報といった統計的な情報を用いるものが一般的である．

しかし，検索意図に適合する文書の弁別を，出現単語の統計量に現れた特徴のみで行うことには限界があるため，近年では，自然言語処理技術を用いて，文書に現れた構文的／意味的な情報（言語的情報）を抽出し，その情報を類似度計算にとり入れて，文書ランキングを高精度化しようとするアプローチが提案されている[2]．

我々は，従来の統計的なアプローチに加え，文書の構造情報を用いて重要部分を判別し，さらに，その重要部分での単語の句内共起表現を抽出して利用することにより，文書ランキングを高精度化する手法を考案した．本稿では，その手法と，それを特許文書検索に対して適用した実験結果について述べる．

2 ランキング手法

本手法の基本的アイデアを以下に述べる．

・**共起情報の利用** 検索意図あるいは文書内容を忠実に表現する言語的情報として句内共起表現を用いる．検索意図や文書内容は，単一語よりも，複合語や句などの長い構文的機能単位で表現されている場合に，より限定的な概念となり，明瞭に弁別できると考えるからである．

・**文書からの共起情報の抽出** 文書の構造に注目し，まず，各文書の内容を特徴的に表している部分をとり出し，次に，この重要部分に現れた句内共起表現のみを**文書共起情報**として抽出することとする．文書の内容を表す重要な句内共起表現は，文書の特定の部分に偏って出現することが多いと考えるからである．

・**検索式からの共起情報の抽出** 入力が検索式で与えられた場合，以下の単語対を**入力共起情報**（C と略す）として抽出することとする．

(a) 複合語の構成語（出現順）
例. (リレーレンズ＋液晶プロジェクタ)
$C = \{$(リレー, レンズ), (液晶, プロジェクタ)$\}$

(b) 論理積演算子で結ばれた単語対
例. (文書＋テキスト) * (検索＋サーチ)
$C = \{$(文書, 検索), (検索, 文書), (文書, サーチ), $\cdots\}$

入力共起情報は文書共起情報との照合に用いるが，これらの単語対は，適合する文書の重要部分においても，句内共起表現として現れやすいと考えるからである．

・**照合方法** 文書共起情報と入力共起情報の類似度により，文書全体をあらい層状にあらかじめ弁別しておき，弁別された各層の文書について，従来の統計情報によるランキングを行うこととする．単語の頻度や分布などの統計情報よりも，言語的情報（共起情報）を重視したランキングを行うことで，精度の向上をはかる．

・**完全一致モデルとの併用** 通常のキーワードからなる論理式を入力とし，完全一致検索を行って求めた文書集合に対して，非完全一致モデルで算出したランキングを出力する．完全一致検索をとり入れることで，利用者が，検索式を操作して検索結果の母集団を明示的に修正することが可能になり，検索もれの心配を軽減することができる．

以上の基本的アイデアに基づき，本手法では以下のアプローチをとる．

(a) 検索質問はキーワードからなる論理式とする．
(b) 文書および検索質問を $(\vec{v}, S)$ の組で表現する．
$\vec{v}$：単語頻度情報から構成したベクトル表現
S：文書の重要部分と検索式から得た句内共起表現
(c) 検索質問から完全一致モデルによって求めた検索結果に対して，ランキングを行う．その際に，(b) の特徴表現を以下のように用いる．
(c-1)：句内共起表現（入力共起情報と文書共起情報）の類似度により，文書を層状に弁別する．
(c-2)：(c-1) で弁別された各層について，検索式から構成した**入力ベクトル**と文書から構成した**文書ベクトル**を用いて，検索式と文書の類似度を比較し，ランキングを行う（**層別ランキング**）．

3 実験方法

特許公報を対象に本手法によるランキングの実験を行った．完全一致モデルは，社内の特許検索システムを利用した．実験用データは，上記システムによる完全一致検索（2回）の検索結果のうち，平成5年以降に公開された特許の明細書全文を利用し，これらの中から関連特許を目視で選別した．

実験	データ規模	うち関連特許
実験1	760件 (13Mb)	85件 (11.2%)
実験2	2,559件 (63Mb)	89件 (3.5%)

・検索式 検索式は，論理和標準形（conjunctive normal form）または論理積標準形（disjunctive normal form）のいずれかの形で記述されており，かつ，演算子に論理積または論理和のいずれかのみを用いているものに限定した．

・特徴表現 入力ベクトルは，検索式で指定されたキーワードまたはその構成語で辞書にある語を索引語として取り出し，各々の索引語の重みを1として構成した．一方，文書ベクトルは，単語の文書内の頻度と文書間の分布を全文データについて算出し，各文書毎に $tf \cdot idf$ による単語の重みづけ[3, page 63]を行って構成した．

また，文書の重要部分として，特許明細のうち，特許

A Ranking Strategy Incorpolating Document Structure and Cooccurrence
Masako Nomoto(E-mail:nomoto@trl.mei.co.jp),
Naohiko Noguchi
Matsushita Electric Industrial, Co., Ltd.
5-15, 4-chome, Higashi-Shinagawa,
Shinagawa-ku, Tokyo 140 Japan

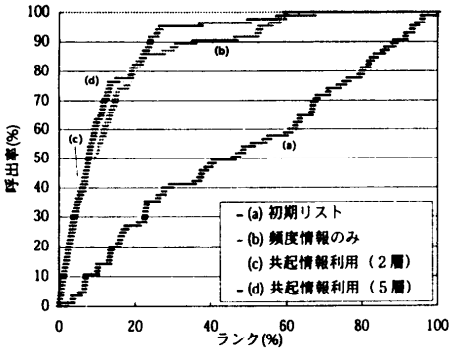

図 1: 実験 1: ランクと呼び出し率の関係

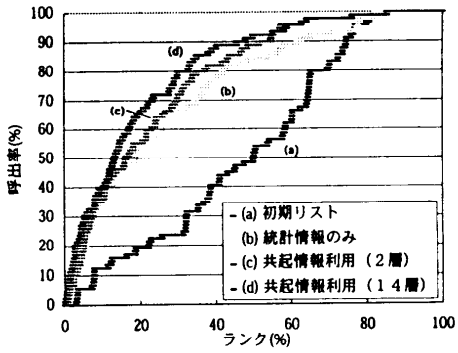

図 2: 実験 2: ランクと呼び出し率の関係

の関連性を判断するときに有用と思われる，(a) 発明の名称，(b) 産業上の利用分野に関する記述，(c) 従来の技術に関する記述を抽出した[1]．さらに，文書共起情報として，上記の重要部分から，以下の共起関係をもつ単語対を抽出した．

(1) の関係　(名詞)+ 助詞「の」 +(名詞)
　例. (液晶パネル) の (カラーフィルタ基板)
(2) 名詞連続　(名詞)+(名詞)
　例. (偏光)(装置)
(3) 格関係　(名詞)+ 助詞「が / を / に」 +(用言)
　例. (偏光板)+ 「を」 +(分離し)

・照合方法　共起情報の有用性を検証するため，以下の 4 通りのランキングを行った．

(a) 初期リスト: 完全一致検索の結果 (日付順)
(b) 統計情報のみによるランキング
(c) 統計情報 + 共起情報による層別ランキング (二層)
(d) 統計情報 + 共起情報による層別ランキング (多層)

文書の階層化の方法として，(c) では文書共起情報が入力共起情報を含むかどうかで二層に分ける方法，(d) では文書共起情報が何種類の入力共起情報を含むかで多層に分ける方法をとった．また，(b)，(c)，(d) で統計情報によるランキングを行う際には，入力ベクトルと文書ベクトルの類似度を測る尺度として，ベクトルの内積尺度を採用した．

4 結果

ランキングと関連特許の呼び出し率の関係を図 1, 図 2 に示す．(a)，(b)，(c)，(d) のパフォーマンスを正規化呼出率 (normalized recall[3, page 181] : 以下，NR と略す) により比較した結果，および (a) を基準としたときの各々の改善率は以下の通りである．

実験 1	(a)	(b)	(c)	(d)
NR	0.536	0.906	0.937	0.938
改善率	1	1.69	1.75	1.75

実験 2	(a)	(b)	(c)	(d)
NR	0.541	0.756	0.791	0.822
改善率	1	1.40	1.46	1.52

[1]特許の明細書では，「発明の詳細な説明」の項目で産業上の利用分野および従来の技術に関して記述されることが多いが，これらの記述は必須ではない．今回の実験で，文書の重要部分として，これらの記述の抽出を試みたところ，各々約 94% の特許明細から抽出できた．

実験 1，2 とも，各ランキングの NR は上位から (d), (c), (b) の順であり，統計情報のみよりは共起情報を加えた方がよく，共起情報による層分けは二層より多層にした方が効果的であることが分かる．実験 1 では (b), (c), (d) の差が小さいが，(c), (d) で上位の非関連特許では，入力共起情報と同じ共起表現 (例. 「プローブ顕微鏡」) が文書共起情報に含まれており，これらの非関連特許が層別ランキングで上層にランクされた結果，(c), (d) の精度がやや落ちたと考えられる．

5 結び

従来の統計情報によるランキングに，共起情報によるあらい層状の順位付けを加えてランキングを行う手法を提案した．

今回の実験を通じて，層別ランキングは，統計情報のみによるランキングよりも，すぐれたパフォーマンスを示した．また，共起情報の種類で多層に分けてランキングを行うとより効果が上がることから，層別ランキングの有効性を確認した．さらに高精度化するために，今後の課題として，以下の点が挙げられる．

・入力共起表現の重み付けによる多層化
・重要部分の選択箇所のチューニング
・ユーザからのフィードバック情報の利用

今後はこれらの課題を解決しながら，システムの実用化に取り組む予定である．

謝辞

当研究に関して特許調査データを提供していただくとともに，有益なコメントをいただいた当社知的財産権センターならびに松下技術情報サービス (株) の方々に感謝致します．

参考文献

[1] Belkin,N.J. and Croft, W.B. *Retrieval techniques*, Annual Review of Information Science and Technology, Vol.22. Williams, M.E., ed. Amsterdam, Elseview science Publishers, 1987, pp.109-145.

[2] Smeaton,A.F. *Progress in the Application of Natural Language Processing to Information Retrieval Tasks*, The Computer Journal, Vol.35. No.3, 1992, pp.268-277.

[3] Salton, G. and McGill, M.J. *Introduction to Modern Information Retrieval.* New York, McGraw-Hill, 1983.