3N－6

極大単語索引を用いた知的検索ソフトウェア *MEISTER* －大規模文書検索への応用－

野本昌子　野口直彦　菅野祐司　佐藤光弘　稲葉光昭　福重貴雄
{nomoto,noguchi,kanno,msato,inaba,fuku}@trl.mei.co.jp
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

筆者らが開発中の知的検索ソフトウェア *MEISTER*[1] は，索引の複数化，検索処理の並列化等により，新聞記事，特許などの大規模文書検索に対しても適用可能である．本稿では，*MEISTER* を用いて実現した特許検索の実験システムの概要と，検索精度および関連キーワード機能の評価について述べる．

2　特許検索実験システムの概要

本システムは，高精度な文書ランキングにより，特許文書を効率よく検索する．大規模なデータを検索するため，索引の複数化，検索処理の並列化を行ない，現在，特許明細書４年分を対象に所内で実験運用を行なっている．実験システムの構成を図 1に，また，本システムの特長，実験運用の状況を以下に示す．

(a) **大規模文書検索**: 現在，特許明細書データ４年分 (1,375,720件，約 28GB) を対象にした検索が可能．

表１　実験システムの規模

辞書 (43万語)	拡張位置索引 (圧縮)	語リスト索引	共起索引	表示用データ
26.5MB	21GB	2.5GB	2.1GB	5.6GB

(b) **高精度なランキング**: 統計情報に加えて文書構造と共起情報を用いた高精度な文書ランキング[3]を実現．

(c) **柔軟な検索インターフェイス**: 文章または論理式による条件指定および単語検索，文字列検索が可能．

(d) **関連キーワード機能**: 関連キーワードの提示と，関連キーワードを検索条件に追加した再検索が可能．

表２　実験システムの運用状況 (中間結果)

期間	1997年２月～７月
質問数	計 1,419 質問 (文章型:764, 論理式型:655)
検索条件	(文章型) 平均 8.2 文字,(論理式型) 平均 2.9 語

3　実験 1: 検索精度の評価

本システムのランキングの精度を測定する実験を行なった．まず，社内の部署で実際に特許調査に用いた複雑な論理式 (3種類) により，特許明細書全文 (４年分) を対象に，従来型の特許検索システムで検索を行ない，各々の検索結果の文書集合から関連特許を目視で選別して，以下の評価セットを作成した．

表３　評価セット

評価セット	検索結果	うち関連特許
(α)	1,388件	350件　(25.2%)
(β)	520件	61件　(11.7%)
(γ)	2,559件	89件　(3.5%)

Maximal-Extension Indexing method for Smart TExt Retrieval
MEISTER:Application to large document retrieval
Masako Nomoto, Naohiko Noguchi, Yuji Kanno,
Mitsuhiro Sato, Mitsuaki Inaba, Yoshio Fukushige
Multimedia Systems Research Laboratory,Matsushita Electric Industrial, Co., Ltd.
4-5-15, Higashi-Shinagawa,Shinagawa-ku,Tokyo 140 Japan

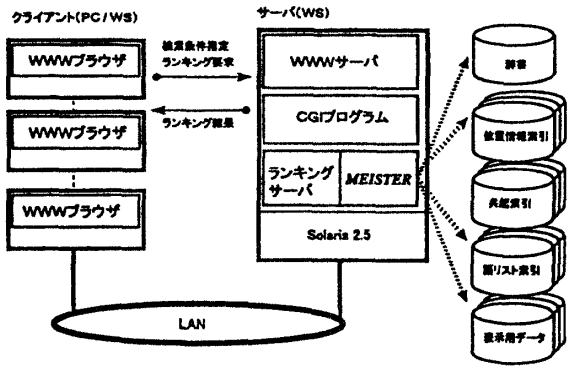

図１　実験システムの構成

その後，各評価セットに対して，本システムによるランキングを行なった．本システムでは *MEISTER* のランキングライブラリ[4]を用いており，各論理式に含まれる単語集合を $Q = \{q_1, \cdots, q_n\}$ とした時，文書 Doc_j と Q の関連度 $Rel(Q, Doc_j)$ は，以下のように３種類の評価基準の組み合わせにより表される．

$$Rel(Q, Doc_j) = C_1 \sum_{q_i \in Q} w_i \cdot tf_{ij} \cdot idf_i + C_2 M_j + C_3 N_j$$

ただし，C_1, C_2, C_3 は定数，w_i は単語 $_i$ の重み，M_j，N_j は各々，文書内共起の度合[2]，構文要素内共起の度合[2]を表す．

今回の実験では，$w_i = 1$ とし，M_j は，Q 中の単語のうち Doc_j 内に出現したものの数，N_j は，論理式から自動抽出した入力共起情報 (論理積演算子で結ばれた単語対および複合語を構成する単語対) のうち，文書の重要部分 (発明の名称，出願人，要約，特許請求の範囲，発明の詳細な説明の最初の 10 行) に構文要素内共起 (の関係，名詞連続，格関係)[3]として出現したものの数とした．また，$tf_{ij} \cdot idf_i$ は q_i の文字列としての出現頻度および分布を用いた．

今回は，上記の３つの評価基準を，構文要素内共起 (N_j)，文書内共起 (M_j)，$tf_{ij} \cdot idf_i$ の順に優先して組み合わせ，従来の統計情報 ($tf_{ij} \cdot idf_i$) のみによるランキングと精度を比較した．また，Doc_j 内の構文要素内共起を評価する際，論理式から自動抽出した入力共起情報とそのまま照合する場合と，人手で修正した入力共起情報と照合する場合との比較も行なった．以下の (a) ～ (d) の設定でのランキングの適合率，再現率を，上記の３つの評価データで平均した結果を図 2に示す．なお，グラフは，再現率の値 0.05 毎に平滑化した．

(a) $C_1 = C_2 = C_3 = 0$ (日付順の初期リスト)

(b) $C_3 >> C_2 >> C_1$ (手法 1: 入力共起情報を全て利用)

(c) $C_3 >> C_2 >> C_1$ (手法 2: 入力共起情報を修正)

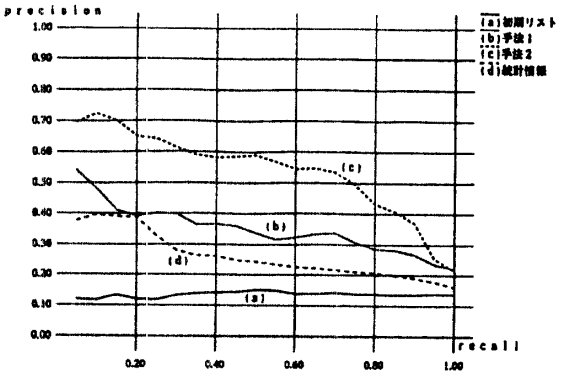

図2　実験1: 検索精度の比較

(d) $C_2 = C_3 = 0$　　(統計情報のみ)

本手法 (b), (c) の精度は統計情報のみを用いたランキング (d) を上回った. さらに, 論理式から抽出した全ての入力共起情報を用いる (b) に対して, (c) のように人手で共起に修正を加えることで, 大幅に精度が向上することを確認した.

4　実験 2: 関連キーワード機能の評価

関連キーワード機能の効果を確認するため, 関連キーワードの一部を元の検索条件に追加して再検索を行ない, 精度が向上するかどうかを調べた. 今回の実験では, 実験1の3つの評価セットのうち, (γ) を用い, ランキングの設定は, 前記の3つの評価基準のうち, 文書内共起 (M_j), $tf_{ij} \cdot idf_i$ の2つをこの順に優先することとした ($C_3 = 0, C_2 >> C_1$). 以上の設定による検索結果に, 以下の (a) ～ (c) の方法で関連キーワードの追加と再検索を4回繰り返し, 計5回の検索を行なった.

(a) ランキング上位10文書から抽出した関連キーワード (上位50) から, 利用者が目視で選択.

(b) ランキング上位10文書から抽出した関連キーワード (上位50) のうち, (a) と同数のキーワードを, 上位から機械的に選択.

(c) ランキング上位の関連特許10文書から抽出した関連キーワード (上位50) のうち, 上位 (1 ～ 10位) のキーワードを機械的に選択.

各々の検索結果における, 上位10位までの適合率を以下に示す.

表4　上位10位までの適合率

方法	1回	2回	3回	4回	5回
(a)	0.0	0.0	0.2	0.2	0.3
(b)	0.0	0.1	0.1	0.1	0.1
(c)	0.0	0.4	0.6	0.4	0.4

(a), (b), (c) とも, 最初の検索では上位10位までに関連特許はなかったが, 関連キーワードを元の検索条件に追加して再検索を繰り返すことで, 10位までに関連特許が現れている.

また, 最初の検索結果と (a), (b), (c) の2回目および5回目の検索結果におけるランキングと再現率の関係を各々図3, 図4に示す.

図3の (a), (b) より, ランキング上位文書から抽出した関連キーワードを検索条件に追加して再検索を行なうことで精度を向上していることが分かる. 特に今回の実験のように, 最初の検索結果でランキング上位10件中に関連文書がない場合でも, 関連キーワードのうち適切なものを直接人手で選択する (a) は有効であり, 再検索を繰り返すとさらに効果が上がる. また, これらの関連キーワードの選択による再検索は, 通常の適合性フィードバックの手法に比べて, 文書を読んで関連文書を選択する手間を軽減できるメリットがある.

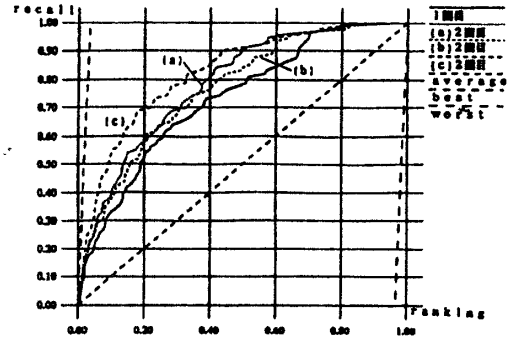

図3　実験2-1: ランキングと再現率の関係

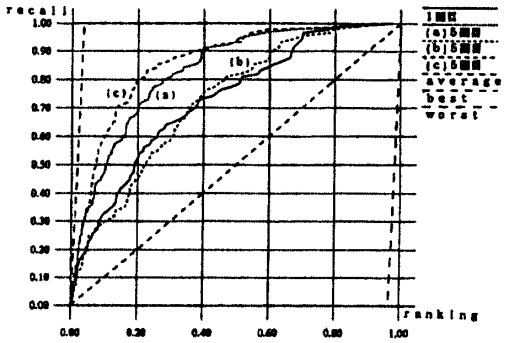

図4　実験2-2: ランキングと再現率の関係

一方, (c) は適合性フィードバックにより, 関連文書のみから関連キーワードを抽出して関連キーワードの精度を高めるものであり, 図3, 図4に見られる通り, (a), (b) のように関連キーワードを直接選択するより, さらに有効である. また, 今回 (c) では関連キーワードを上位から機械的に条件に追加したが, これに (a) を組み合わせて, 適切な関連キーワードを人手で選択するようにすると, 再検索の一層の精度向上が期待できる.

5　結び

知的検索ソフトウェア *MEISTER* を応用した特許検索実験システムの概要とランキング精度および関連キーワード機能の評価結果について述べた.

今後は, さらに大規模化したデータで実験を重ね, 検索速度の高速化, 書誌事項検索, 個々の単語や共起情報のランキングにおける影響力の提示などの課題を解決して, さらに有効な特許検索システムの構築をめざす.

参考文献

[1] 野口直彦, 他: 極大単語索引を用いた知的検索ソフトウェア *MEISTER*- 概要 -, 第55回情処全大, 3N-1(1997).

[2] 野口直彦, 他: 単語統計情報と言語情報とを併用した新しい文書検索のモデル, 情処研報, 96-FI-44-5(1996).

[3] 野本昌子, 野口直彦: 文書構造と共起表現を用いた文書ランキング手法, 第52回情処全大, 5P-6(1996).

[4] 稲葉光昭, 他: 極大単語索引を用いた知的検索ソフトウェア *MEISTER*- ランキングライブラリの機能と特長 -, 第55回情処全大, 3N-3(1997).