3N－4

極大単語索引を用いた知的検索ソフトウェア *MEISTER* — 関連キーワードライブラリの機能と特長 —

佐藤光弘　野口直彦　菅野祐司　野本昌子　稲葉光昭　福重貴雄
{msato,noguchi,kanno,nomoto,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

電子化文書を検索する際，ユーザは自分の欲しい情報を手に入れるために検索条件を様々に試行錯誤することが多い．しかしながら，対象文書群や必要とする文書の特性に応じた適切なキーワードを想起することは難しく，検索の効率を落とす一因となっている．知的検索ソフトウェア *MEISTER*[1]の関連キーワードライブラリは，上記の問題を解決すべく開発したものであり，検索結果の絞り込みや関連トピックへの移動といったユーザの再検索支援に有効な関連キーワードを自動抽出する機能を備えている．本稿では，関連キーワードライブラリの機能，および性能評価の実験結果について報告する．

2　機能概要

関連キーワードライブラリは，何らかの手順で特定された文書集合 (文書番号リスト) を入力とし，その文書集合に特徴的な単語を関連キーワードとして抽出する．図1に関連キーワード抽出の基本原理を示す．

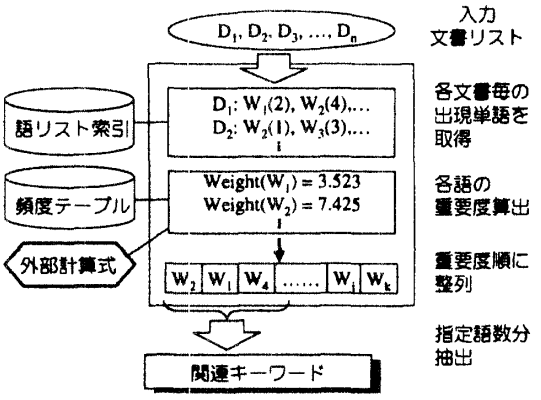

図1　関連キーワード抽出の基本原理

文書リストから出現単語とその出現回数を得るために，辞書 / 索引ライブラリの語リスト索引を利用する．各語の重要度計算は，$tf \cdot idf$ に文書リスト中での出現頻度を加味した値を基本とし，パラメータ設定により数種類の正規化手法が利用できる．また，各語に重みを付与する計算式を外部から与えることも可能である．

各語の頻度情報取得には，辞書 / 索引ライブラリにより作成された頻度／拡張位置索引から各語の出現文書数と総出現頻度のみを格納した頻度テーブル (本ライブラリで提供するツールにより作成) を利用する．

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

さらに本ライブラリでは，上記の基本原理に加えて，以下に示す除外語設定が可能である．

(1) 高・低頻度語
高頻度語は特定文書集合中でも頻出する可能性が高いが重要度は低い．また，例えば１文書にしか出現しない語は関連キーワードとしての利用価値が低い．これらを考慮して，本ライブラリでは，全文書数に対する出現文書数の上限～下限 (パラメータにより設定) の範囲に含まれない出現頻度を持つ語は除外する．

(2) 延長語－部分語
抽出されたキーワード集合の任意の２要素に延長語－部分語の関係があると，関連キーワード全体として冗長になる．本ライブラリでは，単語間に部分語関係がある場合，与えられた除外規則に応じていずれかを除外することができる．除外規則は，(1) 部分語を除外 (2) 延長語を除外 (3) 重要度の低い語を除外 の３種を実装している．

(3) 外部計算式による指定
外部から指定する計算式の中で，特定の語に対して重み０を返す計算式を設定すると，その語を除外する．例えば，辞書を利用して，キーワードとして的確でない品詞を持つ語（自立語でないものなど）は除外する，といった計算式を定義できる．

3　実験１

本ライブラリにより抽出されたキーワードの精度を評価する実験を行った．

まず，特許の明細書全文 (４年分) に対し，社内の特許検索システムにより３種類の検索式による検索を行い，各々の検索結果文書の中から関連特許を目視で選別し，これを正解文書集合とした．さらに，各々の関連特許から無作為に10件の関連特許を選び，これに対するPATOLISキーワードを取得した．

こうして得られたPATOLISキーワードを，キーワード抽出の正解集合と仮定する．ただし，PATOLISキーワードのうち，一文字語，アルファベットのみの語，特許明細書で一般的な語 (“装置”，“方式” など) は除外した．表１に，各評価セットごとの数値を示す．

表１　実験で利用した評価セット

評価セット	検索結果	うち関連特許	10関連特許 PATOLISキーワード数
(1)	1,388件	350件	240語
(2)	520件	61件	257語
(3)	2,559件	89件	239語

実験は，PATOLIS キーワード取得に利用した 10 文書を関連キーワードライブラリに与えてキーワード群を抽出し，これらの PATOLIS キーワードに対する適合度を算出した．関連キーワードは，ランク上位 50 語とし，一文字語・アルファベット語・名詞以外の品詞を持つ語は除外するように外部計算式を設定した．

これを 3 種類の評価セットに対して実行し，得られた適合率の平均をとった．図 2に適合率のグラフを示す．

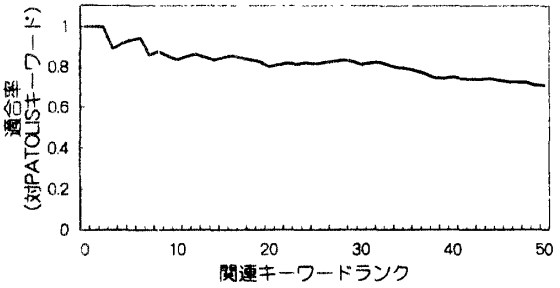

図 2　関連キーワードの適合率

実験の結果，上位 30 語までの約 8 割が PATOLIS キーワードと一致しており，抽出された関連キーワードが非常に高精度であることが確認できた．

4　実験 2

PATOLIS キーワードの中には，その文書に特徴的である語とより一般的な語とが混在しているため，関連キーワードとして得られた語の再検索における有効性は，PATOLIS キーワードに対する適合率だけは判断できない．そこで，関連キーワードとして抽出された語を再検索に利用した場合の効果を確認するため，次の実験を行った．

1. 実験 1 と同様の条件で，関連キーワードライブラリにより 500 語を抽出
2. 抽出した関連キーワードに含まれる PATOLIS キーワードを，その重要度の順に整列したリストを作成
3. 上記に含まれなかった PATOLIS キーワードを文字コード順に整列し，リストの最後尾に結合
4. 特許検索に利用した検索式に，上記リストの各キーワードを個別に OR 条件で追加し，ランキングを実行して得られた検索結果の正解文書集合に対する Normarized Recall[2] (以下 NR と略す) を算出

追加したキーワードが正解文書集合に特徴的なものであれば，NR が向上することが予想できる．

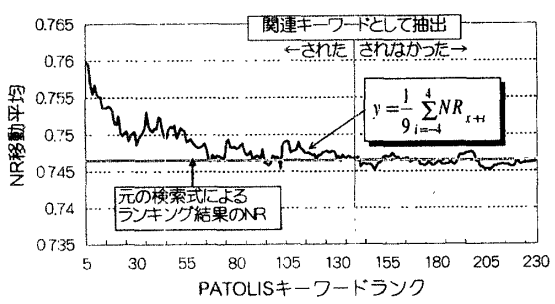

図 3　PATOLIS キーワードランク − NR 曲線

(1) の評価セットについて，ランキングした PATOLIS キーワードの NR の推移を図 3に示す．値にばらつきがあるため，図中に示す計算式により移動平均をとった値をグラフ化した．また図中央の縦線付近までが，本ライブラリで抽出できた語である．

グラフから，NR を向上させるキーワード (すなわち正解文書集合をより特徴づける語) が上位に集中しており，PATOLIS キーワードの中でも，再検索時に有効性の高い語が関連キーワードライブラリによって抽出されていることが確認できた．

5　考察

実験 1 の条件での関連キーワード抽出速度は，PanaStation SS-UA2 (UltraSPARC-I × 2, 200MHz, 主記憶: 256MB) 上で平均 0.3 秒程度と，実用上問題ない速度であった．また，実験 1 および 2 により，本ライブラリの関連キーワード抽出精度も高いことが確認できた．

本ライブラリを応用した検索システム構築の際には，以下の実装が考えられる．

(1) ランキングライブラリとの連携
MEISTER のランキングライブラリで得られるランキング結果の上位文書群は，入力された検索式に対する関連度が高いことが期待できる．そこで，この上位文書群を関連キーワードライブラリの入力とすることで，検索式に対する関連度が高く，かつ実文書の特性に合った関連キーワードを得ることが可能となる．このような実装を行った実システムとして，当社で開発したホームページ知的検索システムがあり，絞り込みだけでなく，トピックを移動するような連想型検索にも関連キーワードが有効であることを確認している[3]．

(2) 適合性フィードバック
本稿における実験は，主に抽出された関連キーワードを再検索に利用する際の有効性を確認するものであり，関連キーワード抽出に利用した文書集合は関連文書であることを人手でチェックしたものの一部であった．したがって，検索結果から関連文書を目視により選択し，ここから抽出される関連キーワード群を利用して質問拡張を行なう (適合性フィードバック) という実装でかなり有効に機能することが予想できる．実際に特許文書を対象とした実験システムでその有効性が確認されている[4]．

今後は，関連キーワード抽出の精度向上を図るとともに，上記システム以外にも本ライブラリを応用し，評価・改良を行っていく予定である．

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] Salton,G. and McGill,M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill Publishing Company (1983).

[3] 佐藤光弘 他: 極大単語索引を用いた知的検索ソフトウェア *MEISTER* – ホームページ検索への応用 –, 第 55 回情処全大, 3N-7 (1997).

[4] 野本昌子 他: 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 –, 第 55 回情処全大, 3N-6 (1997)