3N－3

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 —

稲葉光昭　野口直彦　菅野祐司　佐藤光弘　野本昌子　福重貴雄

{inaba,noguchi,kanno,msato,nomoto,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

大量の電子化文書を検索する機会が増え，検索結果を絞り込むために，高速に文字列検索を行うだけではなく，関連度評価に基づいて文書をランキングしたり，検索ノイズを減らすための工夫が必要となってきている．

著者らが開発中の知的検索ソフトウェア *MEISTER*[1]のランキングライブラリは，上記のような問題点を解決するために論理演算／ランキング／単語検索機能を備えている．本稿ではこれらの機能と実現方式，および性能評価の実験結果について報告する．

2　機能の概要

ランキングライブラリは，以下のような豊富な機能を持ち，これらの組み合わせにより，様々な検索の形態に対応することが可能である．

(1) 検索条件入力モード

- **文字列モード** 複数の文字列を羅列して入力する
- **文章モード** 日本語の文章を入力する
- **論理式モード** 論理式を入力する

文字列モードでは，入力された複数の文字列をORで検索する．文章モードでは，入力文から辞書単語を切り出し，それらをORで検索する．論理式モードでは，論理式にしたがって検索する．論理式モードで使用できる演算子には，論理和(OR)，論理積(AND)，否定(NOT)の他に，頻度情報付与(ADD)演算子がある．

共起条件として単語の順序対の入力も可能である．

(2) ランキング機能

ランキング機能は，検索条件との関連度によって文書を並べ替えた結果を提示する機能である．ランキングライブラリでは，通常の $tf \cdot idf$ 重み付けによる評価基準[2]の他，単語共起に基く評価基準も採り入れ，以下のような3種類の評価基準の線形結合により検索条件 $Q = \{q_1, \cdots, q_n\}$ と文書 Doc_j の関連度 $Rel(Q, Doc_j)$ を求め，ランキングを行う．

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

ただし，w_i はターム q_i に与える重み，tf_{ij} は Doc_j における q_i の出現頻度，idf_i は q_i の inverted document frequency，M_j は文書内共起[3]の度合，N_j は構文要素内共起[3]の度合，C_1, C_2, C_3 は定数である．

(3) 単語検索機能

本ライブラリでは，文字列検索における検索ノイズを除去するために単語検索機能を提供している．単語検索機能には，排除する語の種類により以下の3つのモードがある他，排除する語を個々に指定することもできる．

- **完全一致単語検索**「グラフ」で「パラグラフ」「グラファイト」「シリコングラフィックス」を排除
- **前方一致単語検索**「トップ」で「トップ会談」は検索，「ストップ」「ストップウォッチ」を排除
- **後方一致単語検索**「カルテ」で「電子カルテ」は検索，「カルテット」「価格カルテル」を排除

3　実現方式

3.1　頻度ストリーム

辞書／索引ライブラリ[4]では，検索の中間結果を図1に示すような，文書番号と頻度情報(およびスコア)を両方持つ，頻度ストリームと呼ぶ表現形式で統一的に扱っている．このため，ランキングのスコア計算に必要な情報を蓄積しつつ，効率よく論理演算を行うことができる．

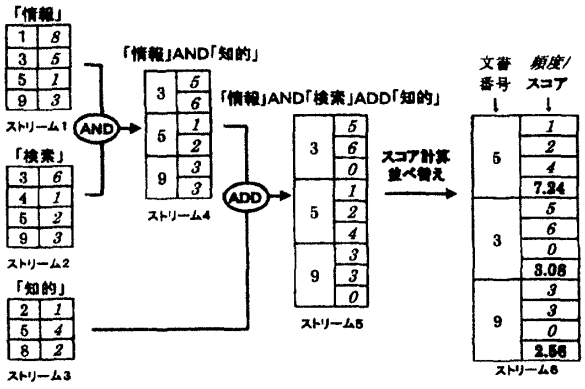

図1　頻度ストリーム

図の例は検索条件式((情報 AND 検索)ADD 知的)を与え，論理検索とランキングを行った場合である．

まず，個々の検索文字列で索引を引き，それぞれの頻度ストリーム1～3を得る．続いて，「情報」と「検索」の論理積(AND)をとる(ストリーム4)．結果ストリームの頻度情報にはストリーム1，ストリーム2の両方の頻度情報が格納される．

次に，論理積の結果に「知的」の頻度情報付与(ADD)を行う(ストリーム5)．頻度情報付与(ADD)は，検索結果の文書集合には影響を与えず，ランキングの順序だけに影響を与えるような検索文字列を指定するために用いる．ストリーム5は，文書集合はストリーム4と同じで，ストリーム3と共通する文書番号5の頻度情報があらたに付加されているのがわかる．

最後に，蓄積した頻度情報をもとに各文書のスコアを算出し，スコアの高い上位Nmax文書を求める．

スコア計算，ランキングを処理の最終段階で一括して行うため，複数の計算機に索引を分割して置く場合に大域的な *idf* の算出を待たずに、ストリーム間演算が行える点や，同時に複数のストリーム間演算を行える点で，ナイーブな方法に比べ効率が良い．

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

3.2 条件付文字列検索

単語検索機能は極大単語索引方式によって作成された(拡張)位置索引から，文字列の出現箇所を選択的に取得することで実現できる．辞書／索引ライブラリでは表1-Aのような条件(複数の指定可能)に合う出現箇所を効率良く選択的に取得できる．

なお，sは検索文字列，Dは辞書，P(X)は索引に登録されている語の集合X(X⊆D)中の各語の照合文字位置の和集合，Er(s),El(s)は，各々sを最左部分文字列，最右部分文字列として持ち，かつsでないD中の全ての語，Eb(s)はsを部分文字列として持ち，かつsでもEr(s)でもEl(s)でもないD中の全ての語を表す．

表1 条件付き文字列検索

[1-A]

	条件	取得出現箇所
(1)	MAXIMAL	P({s}) (s∈D) φ (s∉D)
(2)	PREFIX	P(Er(s))
(3)	POSTFIX	P(El(s))
(4)	INFIX	P(Eb(s))
(5)	EXTENDED	(1)+(2)+(3)+(4)
(6)	NON_EXTENDED	(7)−(5)
(7)	ALL(文字列検索)	全出現箇所

[1-B]

検索モード ＼ 検索文字列	辞書単語	非辞書単語
完全一致単語検索	(1)	(6)
前方一致単語検索	(1)+(2)	(6)+(2)
後方一致単語検索	(1)+(3)	(6)+(3)

ランキングライブラリでは，この条件付き文字列検索機能を用いて，与えられた検索文字列が辞書中の語である場合と，そうでない場合に分け，表1-Bのようにして単語検索を実現している．

4 実験

新聞記事本文2.5年分(約77.7万記事,708MB)から作成した(圧縮)拡張位置索引(737MB)を用いて，文字列検索速度／ランキング速度，ランキングの精度の評価を行った．実験にはPanaStation SS-UA2(UltraSPARC-I×2,200MHz, 主記憶:256MB)を使用し，検索時間は各検索条件に対し，2回検索を実行した2回目を実時間で計測した．

実験1，2について、検索条件は「情報検索システム評価用ベンチマーク」の60質問を基に，((製造 AND 販売 AND 一体化)OR 製販一体化) の様な，平均3.5文字列からなる論理検索式を人手で構成して与えた．

4.1 実験1：文字列検索／ランキング速度

図2に60質問に対する文字列検索およびランキング速度の測定結果を示す．求めるスコア上位文書数Nmaxは30とした．1質問あたりの平均検索時間(文字列検索，論理演算，ランキングにかかる時間)は0.4秒程度であり，実用上問題のない検索時間であること，また，そのうち，論理演算とランキングにかかる時間の割合は20%程度であり，文字列検索時間に比べ，充分に短いことが確認できた．

4.2 実験2：ランキング精度

ランキング精度の実験には，数値範囲指定，シソーラス展開が本質的である質問を除いた19質問を与え，検索結果の正解／不正解の判断は人手により行った．図3に各質問に対するランキング結果の上位30件までの平均適合率を示す．レコード番号順に並べた場合に比べ，ランキングを行うことで適合率が2倍程度改善できることが確認された．

4.3 実験3：単語検索

新聞記事0.5年分に対し検索文字列「カルテ」で検索を行い，人手で正誤判定を行った結果，文字列検索では適合率37.8%(正解79/検索結果数207)だったのに対し，後方一致単語検索では92.9%(正解79/検索結果数85)であった．単語検索機能により，検索ノイズが大幅に除去できていることがわかる．

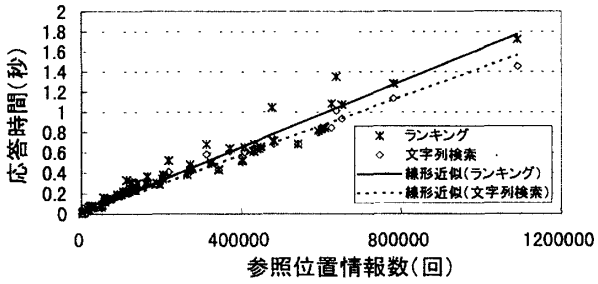

図2 文字列検索／ランキング時間

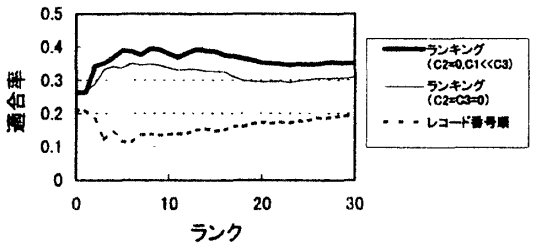

図3 ランキング精度

5 おわりに

知的検索ソフトウェア*MEISTER*のランキングライブラリの機能と実現方式について述べた．

また，新聞記事データを用いた評価実験を行った結果，実用的に充分な検索速度と，ランキングによる精度の向上、単語検索機能による検索ノイズの除去が確認できた．

本研究での実験システムの構築，ならびに評価実験において，日本経済新聞社データバンク局様にご協力をいただきました．ここに深く感謝いたします．

また，評価実験には，株式会社日本経済新聞の協力によって，社団法人情報処理学会・データベースシステム研究会・情報検索システム評価用データベース構築ワーキンググループが，1993年9月1日から12月31日の日本経済新聞記事を基に構築した情報検索評価用データベース(テスト版)を利用させていただきました．

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-01(1997).

[2] 海野敏: 出現頻度に基づく単語重みづけの原理, Library and Information Science, No.26(1988).

[3] 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，情報処理学会情報学基礎研,96-FI-44-5(1996).

[4] 菅野祐司 他: 極大単語索引を用いた知的検索ソフトウェア *MEISTER* - 辞書／索引ライブラリの機能と特長 -, 第55回情処全大, 3N-02(1997).