極大単語索引を用いた知的検索ソフトウェア *MEISTER*
3N－2　— 辞書／索引ライブラリの機能と特長 —

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

筆者らが開発中の知的検索ソフトウェア *MEISTER*[1] の多彩な機能を実現し、幅広い分野に応用するには、高速・高機能で、スケーラビリティー，頑健性，拡張性に優れた、辞書／索引の作成と利用のための基本ソフトウェアが不可欠である。「辞書／索引ライブラリ」は、各種の辞書と索引を作成するツール群と辞書／索引検索用ライブラリから成り、知的検索のための基盤を提供する。本稿ではその機能概略と、特長である高速文字列検索の処理方式、性能評価について報告する。

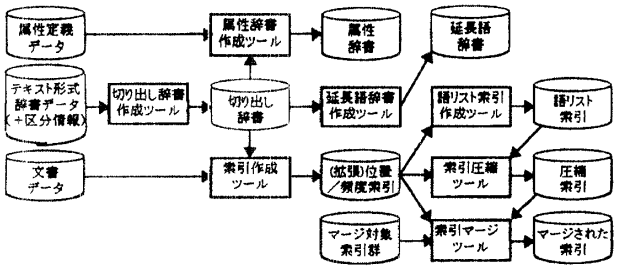

図1　辞書／索引作成の流れ

2 ソフトウェア構成と主な機能

ツールを用いた辞書／索引作成の流れを図1に、ライブラリの主な機能を図2に示す。4種の辞書は利用目的に特化した構造を持つ。特に「切り出し辞書」は Mealy 型のパタン照合機械を含み、文字列から最長一致単語の切り出しを高速に行う。複数の辞書を仮想的に結合して一括検索でき、語彙数／検索機能を拡張できる。索引作成時には、切り出し方式として「完全」「極大」[1] のいずれかを選択できる。高速な文字列検索が必要な場合には、テキスト形式辞書データに後述する「区分情報」を指定して正規表現の切り出し辞書を作り、これを使って正規表現の位置索引（拡張位置索引）を作成する。正規表現辞書／拡張位置索引の検索結果が、単語辞書／位置索引のものと同一となるよう工夫し、一組の辞書／索引で *MEISTER* に必要な機能を提供する。

3 正規表現辞書を用いた高速文字列検索の原理

筆者らが考案した極大単語索引[2]では、単語辞書を用いて「極大切り出し」を行い、「被覆方式」で漏れのない文字列検索を実現した。この方式には、高頻出の単語を含む複合語での検索速度が大幅に低下する問題があった[3]が、今回、正規表現辞書を用いることで、これを解決した。その原理は図3に示す通りで、単語辞書Dの各単語に対し、前後に0個以上の文字クラスの列を付加した1つ以上の正規表現を生成し、正規表現辞書Eを作成する。索引作成時には、このEを用いて正規表現に照合する文字列を切り出し、他の照合文字列から、

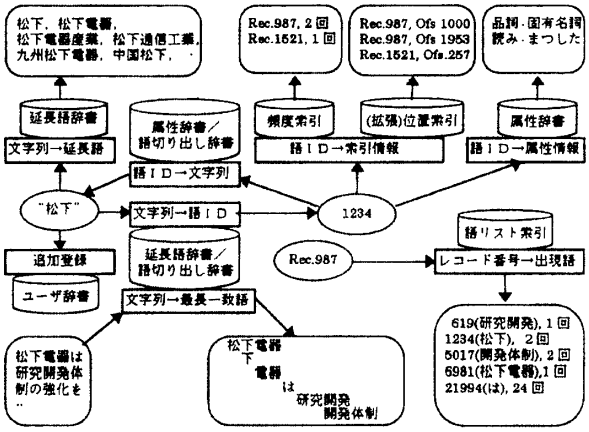

図2　辞書／索引ライブラリの主な機能

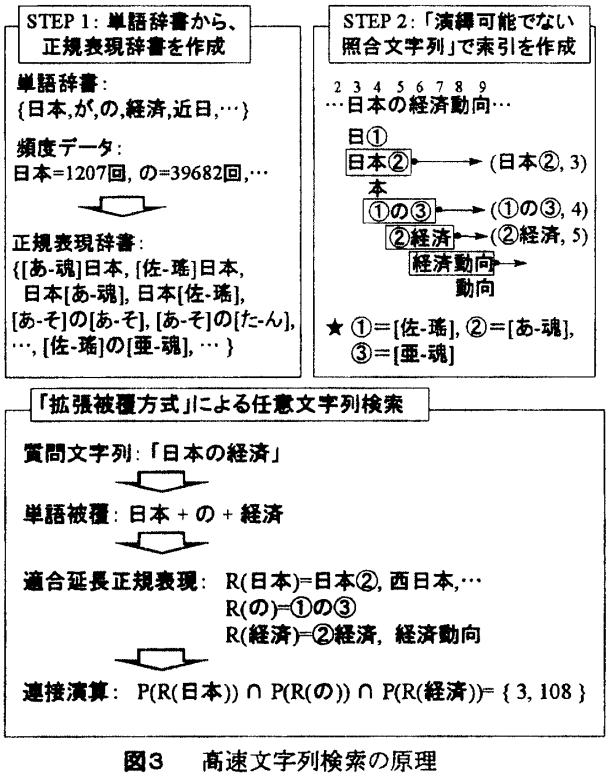

図3　高速文字列検索の原理

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

その文字列の照合が確実に結論されるような「演繹可能な照合文字列」(図の「日①」など)を除き、正規表現と照合文字位置を「拡張位置索引」として登録する。検索時には、まず質問文字列Qの単語被覆(Q全体を覆う、Qの部分文字列であるD中の語Wiの集合)を求める。次に各被覆単語 Wi に対し、Wi を含むEの要素で、Qに適合する(Qを含む文字列に照合する可能性のある)正規表現の集合R(Wi)を求める。最後にR(Wi)の各正規表現の照合文字位置の和集合P(R(Wi))を索引から求め、照合文字位置が連続するものだけを連接演算で取り出し、文字列検索結果を得る。高頻出の単語ほど多数の正規表現を作成することにより、「日本の経済」のように高頻出の語「の」を含む質問の場合でも、連接演算の対象を「①の③」の位置情報だけに絞り込むことができ、検索速度の向上が期待できる。

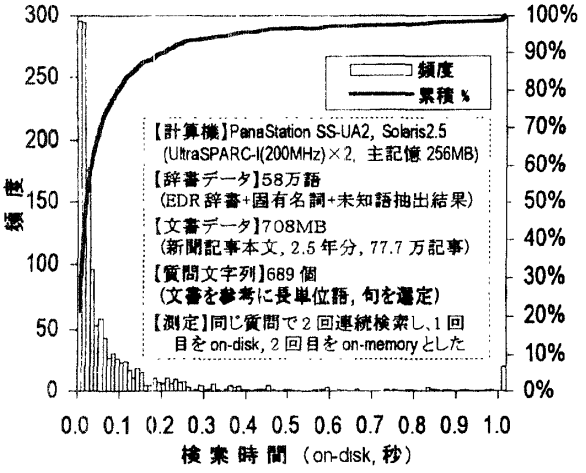

図4 検索時間の分布 (頻度は2回の測定の和)

4 実装・文字列検索性能の評価

上述の原理を辞書／索引ライブラリに実装し、評価した。正規表現は前後に文字クラスを0～1個付加したものを用い、高頻出語ほど作成個数が多くなるよう単語毎に10 段階に区分情報を設定した。文書データを分割して各部分の索引を作成し、検索時に複数の索引を並列検索する機能や、位置情報の上位ビット部分の直前の値との差分を多値算術符号で圧縮する索引圧縮機能も実装した。図4に実験条件および検索時間の分布を、表1に位置索引(単語辞書)との比較結果を示す。拡張位置索引では、高頻出語「の」や「、」を含む質問でも参照位置情報数 $\Sigma|P(R(Wi))|$ が他の質問と同程度に抑えられ、on-disk で数十倍, on-memory で百倍以上の高速化が達成できた。表2は索引圧縮の評価結果で、

表1 位置索引と拡張位置索引の検索性能の比較

質問 (被覆単語数, 関連単語数)	位置索引		拡張位置索引	
	参照位置情報数	検索時間(秒) disk/memory	参照位置情報数	検索時間(秒) disk/memory
「医薬品メーカー」 (1, 4)	760	0.02/0.01	760	0.02/0.00
「コスト格差」 (2, 476)	51055	0.28/0.07	36486	0.33/0.05
「対米輸出依存度」 (3, 56)	163276	0.15/0.13	16524	0.08/0.02
「新分野の開拓」 (4, 96)	10005567	6.50/6.47	48604	0.18/0.05
「白浜町、千倉町、丸山町」 (8, 129)	17702677	11.64/11.38	24611	0.38/0.03
689 個の質問の平均 (15.65.8)	625686	0.49/0.42	13512	0.09/0.02

表2 索引圧縮時の索引容量と検索時間

索引形式	索引容量 (圧縮率)	平均検索時間 (秒)	
		on-disk	on-memory
非圧縮形式 (32bit符号)	1218MB (1.00)	0.103	0.010
標準形式 (16/32bit符号)	944MB (0.77)	0.098	0.011
圧縮形式 (部分算術符号)	702MB (0.58)	0.092	0.018

索引容量は 32 bit固定長記録の場合の 58%と、原文書量程度になり、補助記憶アクセス量の減少で on-disk の検索時間が短縮された。表3, 図5は索引分割の評価結果で、文書データを2分割し、各部分の索引を並列作成することで、作成時間を 3/5 に短縮した。また、異なる補助記憶上の4個の索引を並列検索することで、on-disk の平均検索時間を 3/5 に短縮できた。検索時間が長い質問(図の「不採算製品」, 1索引の検索時間が 0.3 秒)ほど短縮効果が顕著だった。また、同規模の索引4個を並列検索した場合(図の「追加」)の平均検索時間は索引1個の場合の約 1.2 倍だった。この結果は、on-disk の検索時間が on-memory の場合のL倍なので、索引をM分割して並列読出し可能な装置に格納し、N個のCPUの計算機でマルチスレッドによる並行処理を行えば、M＜LNの場合にはMが大きいほど on-disk の検索時間が短くなるためと考えられる。

表3 索引ファイル数と索引作成時間／索引容量

索引ファイル数	作成時間	圧縮時間	合計時間	索引容量
1	42 分 59 秒	6 分 06 秒	49 分 05 秒	702.8 MB
2 (並列処理)	27 分 39 秒 22 分 12 秒	2 分 44 秒 4 分 15 秒	30 分 23 秒 (26 分 27 秒)	707.6 MB

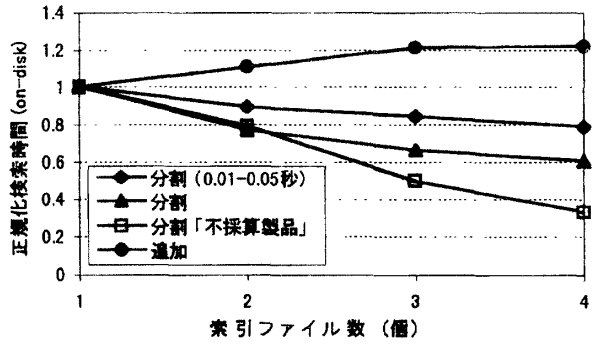

図5 索引の分割と追加による検索時間の変化

5 まとめ

知的検索のための辞書／索引ライブラリの機能と、文字列検索の高速化のための工夫を紹介し、実証規模の実験で検索速度の向上を確認した。正規表現辞書と索引圧縮により、原文書並みの索引容量で実用に耐える文字列検索速度が得られ、索引分割と並列処理によって、大規模化にも対応可能であることを確認できた。

評価実験に使用した新聞記事データを提供して下さった日本経済新聞社データバンク局様に感謝します。

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] 倉知一晃 他: 日本語文書に対する新しい索引検索方式 -索引作成と検索の原理-, 第50回情処全大, 4F-2(1995).
[3] 稲葉光昭 他: 日本語文書に対する新しい索引検索方式 -試作・実験および評価-, 第50回情処全大, 4F-3(1995).