極大単語索引を用いた知的検索ソフトウェア *MEISTER* －概要－

3N－1

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

単語区切りが明瞭でない日本語文書の全文検索において，素朴な単語索引では，辞書の不備／単語分割（形態素解析）精度の限界／未知語の存在等の要因により，任意の文字列については検索漏れが避けられないことから，n-gram 索引を用いた全文検索方式の開発が進められている[1][2][3]．n-gram 索引は，文書中に出現する n-gram（n は可変）を出現位置と共に記録したものであり，検索時には，検索文字列を構成する各 n-gram に対応した出現位置の連接演算により，任意文字列に対する全文検索を漏れなく高速に行う．しかし，n-gram 索引方式には，以下のような課題がある．

(1) **索引量／検索速度** n-gram 索引方式では，通常原文書の数倍の索引容量を必要とする．また，検索文字列長に依存して必要な連接演算が増えるため，高頻出の n-gram を多数含むような検索文字列に対しては，高速化が困難である．

(2) **検索ノイズの除去** 任意文字列による全文検索では，一般に検索ノイズが膨大になる．例えば，「グラフ」という文字列で，「グラファイト」を検索してしまう．n-gram 索引方式は，単語という概念を持たないので，この種のノイズは除去できない．

(3) **文書ランキング等，高度な検索機能** 検索結果を利用者の検索ニーズに関連する順に整列する（ランキング）機能を持つ検索システムは，初期の SIRE，SMART[4]等の実験システムの段階から，90 年代に入って実用化局面を迎え，欧米文書に対しては既にいくつかの商用検索エンジンが開発されている．通常，関連度は，文書中の単語頻度情報を基に算出されるが，n-gram 索引方式では，正確な単語頻度が求まらないので，精密な評価を行うことが困難である．さらに，適合性フィードバックなどの高度な検索機能は別途実現しなくてはならない．

筆者らは，日本語文書に対しても，単語を単位とした索引（完全延長極大索引方式，以下，本稿では極大単語索引方式と呼ぶ）を構成することで，コンパクトな索引で，任意文字列に対して漏れのない高速な全文検索が行えることを示した[5][6]．極大単語索引方式は，従来の単語索引方式と n-gram 索引方式の長所を兼ね備えたものであり，さらに，上記課題を解決することが可能である．

今回我々は，その方式を拡張して文字列検索の高速化・索引作成時間の短縮・索引量の軽減を行い，更に単語頻度情報を用いた文書ランキング等の高度検索機能を実現した知的検索ソフトウエア *MEISTER* を開発した．本稿では，極大単語索引方式の原理と特長，および *MEISTER* の構成と諸機能について述べる．

Maximal-Extension Indexing method for Smart TExt Retrieval *MEISTER*: Overview,
Naohiko Noguchi, Yuji Kanno, Mitsuhiro Sato, 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

2　極大単語索引方式 －原理と特長－

何らかの辞書を用い，文書から単語を切り出して索引に登録することを考える．単語索引とは，切り出された単語と，その文書中での出現位置との組（**索引要素**と呼ぶ）を，必要に応じて登録したものであると考えられる．今，文書に出現する全ての単語を切り出して，その出現位置と共に記録したものを**完全単語索引**と呼ぶことにすると，**極大単語索引**は，次のように定義される．

> 完全単語索引中で，延長関係に関して極大となる索引要素のみを，すべて選択して登録した索引のことを，**極大単語索引**という．

図 1 に，極大単語索引の例を示す．

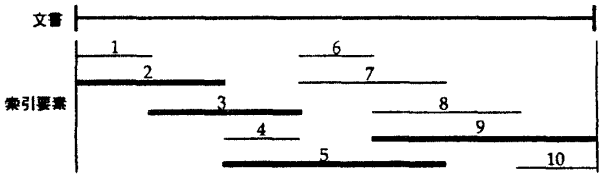

図 1　極大単語索引の例

極大単語索引では，辞書単語の出現位置は，直接索引中に記録されるか，または，その単語の延長語の出現位置として記録されるかの，いずれかとなる．この性質を利用すれば，任意文字列に対する効率的な全文検索アルゴリズムを構築することができる[6]．そのアルゴリズムでは，検索文字列が辞書単語の場合には出現位置の連接演算は必要なく，また，検索文字列が比較的少数個の単語からなる複合語である場合には，文字列自身が長くても，その構成語の出現位置の連接演算のみを行えばよいので，n-gram 索引方式に比べて高速化を図ることができる．また，索引に登録する出現位置数は，文書から切り出される極大単語数であり，文書の全文字位置で情報を記録する n-gram 索引方式に比べて索引容量を軽減することが可能である．さらに，本方式では，極大単語のみの出現位置を登録するので，例えば，「グラフ」という検索文字列に対し，「グラファイト」「グラフィック」など，それを含む延長語の出現位置を検索結果から除去することができ，検索ノイズの軽減が可能である．また，極大単語索引は，単語を単位とした索引なので，文書ランキング・関連キーワード抽出・適合性フィードバックといった高度な検索機能に利用可能である．

本方式と，従来の単語（キーワード）索引方式／n-gram 索引方式との原理上の比較を，表 1 にまとめる．

3　*MEISTER* の構成

知的検索ソフトウェア *MEISTER* は，極大単語索引方式に基づき，2で述べた機能を全て実現したもので

表1 単語（キーワード）索引／n-gram索引／極大単語索引方式の原理上の比較

	機能				性能	
	単語検索	任意文字列検索	ランキング	関連KW	検索速度	索引容量
単語（KW）索引方式	○	×	△	△	◎	◎
n-gram索引方式	×	○	△	×	○	×
極大単語索引方式	○	○	○	○	○~◎	○

ある．現在，本ソフトウェアは，図2に示すようなモジュールからなる，Solàris 2.x 上のライブラリ群として実装されており，それらが提供する諸機能を組み合わせることで，応用システムの要求仕様／規模に合わせ，柔軟なアプリケーション構築を行うことが可能である．

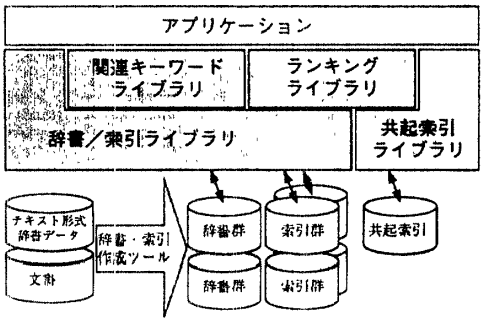

図2 知的検索ソフトウェア *MEISTER* の構成

各ライブラリの機能を以下に述べる．

(1) 辞書／索引ライブラリ

テキスト形式の辞書データから，文書からの単語切り出しを行うための「切り出し辞書」や，単語の属性情報（品詞，接続情報等）を管理する「属性辞書」等の辞書群を構築するツールと，それら辞書群を用いて，単語の各レコード中の頻度情報を登録する「頻度索引」，単語の出現位置情報を登録する「（拡張）位置索引」，レコード毎の出現単語とその頻度を登録する「語リスト索引」等の索引群を高速に構成するツール，および，それら辞書／索引を検索時に利用するためのＡＰＩを提供する．本ライブラリでは，索引に記録された頻度／位置／レコード情報の取得および基本操作関数を提供しており，それらの機能により高速文字列検索を実現している．

(2) ランキングライブラリ

辞書／索引ライブラリのツールを用いて作成された辞書／索引類を利用し，様々な形式の検索条件に対する基本的な検索機能と，関連度評価に基づくランキング機能を実現するための諸関数／ＡＰＩを提供する．本ライブラリでは，論理検索機能（AND,OR,NOT,ADD），単語検索機能（文字列検索のノイズを除去する機能，前方一致／後方一致検索機能も有），ランキング機能（単語頻度情報に基づき，$tf \cdot idf$[4]等の重み付け手法を用いて関連度評価）等の基本機能を実現している．

(3) 関連キーワードライブラリ

関連キーワード機能とは，与えられた文書集合から，その集合内での特徴的なキーワードを抽出する機能であり，検索結果の絞り込み／適合性フィードバック／検索キーワード想起の支援等の際に利用する．本ライブラリは，辞書／索引ライブラリのツールを用いて作成された辞書／索引類を利用し，そのような関連キーワードを抽出するための諸関数／ＡＰＩを提供する．

(4) その他の機能

多言語の文書を同時に検索したいというニーズも高まりつつある．本ソフトウェアでは，辞書部分の拡張を行うことで，英語文書に対する基本的な検索機能にも対応している．現在，完全一致検索機能（単語としての完全一致検索），活用一致検索機能（検索語の活用形に一致する語も検索に含める，規則／不規則活用に対応），複合語検索機能（"White House" などの複合語を検索）等の機能を実現している．

更に，ランキング／関連キーワード抽出等の機能の高精度化のために，単語頻度情報だけでなく，二単語の共起関係の出現情報を抽出して共起索引を構成するツールと，それを検索時に利用するためのＡＰＩを提供する，共起索引ライブラリも用意している．現在，種々の応用システムにて，これらの情報を採り入れたランキング精度の評価実験を行っている[7][8]．

4 応用

本ソフトウェアを用いて，特許／新聞記事など，数ＧＢ～数十ＧＢ程度の大規模文書に対する知的検索システムを開発し，現在，性能／機能評価実験中である．また，近年ニーズが高まっているWWW上のホームページを検索するシステムの開発を行い，（財）地方自治情報センター「地域発見」他，数サイトにて運用が行われている．今後は，本ソフトウェアの評価を行うと同時に，諸機能の高性能化・高機能化を図っていく予定である．

謝辞

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．

参考文献

[1] 赤峯亨 他: 高速全文検索システム RetrievalExpress, 第54回情処全大, 7L-2(1997).
[2] 川口久光 他: n-gram型大規模全文検索方式の開発, 第53回情処全大, 5T-2(1996).
[3] 松井くにお 他: 大容量情報全文検索システム, 1997年電子情報通信学会総合大会, D-4-6(1997).
[4] Salton,G. and McGill,M.J.:*Introduction to Modern Information Retrieval*, McGraw-HillPublishing Company (1983).
[5] 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).
[6] 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).
[7] 野本昌子 他: 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).
[8] 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).