極大単語索引を用いた知的検索ソフトウェア *MEISTER*

3N－7　— ホームページ検索システムへの応用 —

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

近年の World Wide Web(WWW) の急速な普及に伴い，Web ページを対象とする検索システムが脚光を浴びてきている．我々は，知的検索ソフトウェア *MEISTER* [1] を応用し，(1) ネットワークロボットによる特定サイトの網羅的データ収集 (2) 漏れのない文字列検索と高精度な文書ランキング (3) 日本語・英語双方の Web ページ検索が可能 (4) ページの要約文自動生成 (5) 関連キーワード提示による再検索支援 を特長とするホームページ知的検索システム [2] を開発した．本稿では，システムの概要，及び WWW 検索における問題点と本システムが取った解決策について述べる．

2　ホームページ知的検索システムの概要

WWW 検索システムは，データ収集方式に応じて以下のように大別される．

登録型 … 情報発信者・サイト管理者が登録したページ内容の抄録文を検索 (Yahoo 等)

ロボット型 … ネットワークロボットが自動収集したページ全文データを検索 (AltaVista 等)

また，検索対象の範囲についても，広く Web ページ全体を対象とするものと，特定のサイト (例えばある企業グループのサイト群など) のみを対象とするものとがある．本システムはロボット型であるが，検索対象を特定多数の WWW サイト群に限ることで，そのサイト群に関するきめの細かい情報を提供し，特定サイト群全体としての有効な情報発信を支援するシステムとして機能することを目的としている．

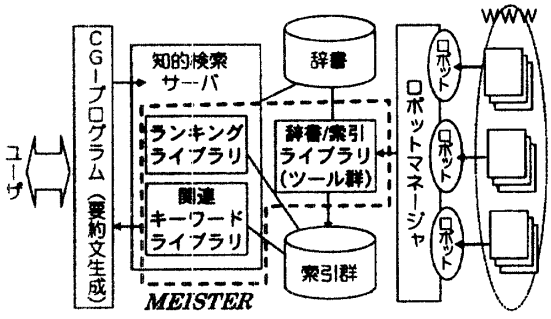

図1　ホームページ知的検索システムの構成

図1に本システムの構成を示す．

Maximal-Extension Indexing method for Smart TExt Retrieval *MEISTER*: Application to WWW search.
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

ネットワークロボットは，いわゆる“ロボット規約”[3] に準拠し，指定された特定サイトのみのデータを収集する．ロボットマネージャは，サイト別に動作するロボットの起動や，収集サイト・収集範囲などの設定を管理するツールであり，これによって，検索サービス運用者の管理にかかる手間を軽減している．

ロボットにより収集されたデータは HTML タグを除去したテキストに変換され，その後，辞書／索引ライブラリのツールにより検索用索引群が作成される．

本システムの知的検索機能は，検索サーバとして実装されており，実際の検索に際しては，検索ページから起動された CGI プログラムが検索サーバと通信して検索結果を得，これを HTML に整形して結果ページを作成しユーザに返す，という処理となっている．検索サーバは，ランキングライブラリと関連キーワードライブラリを利用して構築しており，日本語用・英語用で別サーバとなっている．

3　知的検索機能の特長

図2に本システムによる検索結果の一例を示す．

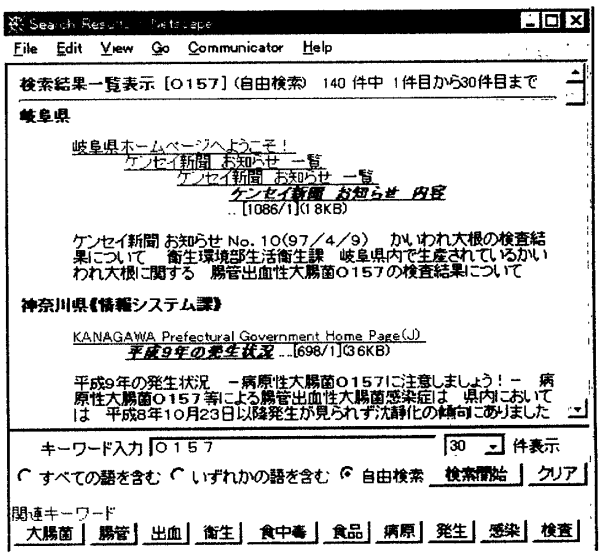

図2　ホームページ知的検索システムの画面イメージ

本システムでは，*MEISTER* の基本機能を応用した文書ランキングと簡単な論理演算をサポートしており，多くの WWW 検索システムと同等の機能を実現している．また，英語ページ検索の場合には，活用一致／完全一致検索を検索時に指定することができる．

WWW 検索では，対象となる Web ページに一般の文書とは異なるいくつかの特徴がある．特に，ネットワークの状態が悪いと実文書を確認するためのコストが非常

に高くなる点，また，ページごとの文書サイズにかなりの格差がある点，などが問題となる．そこで本システムでは，以下の実装を行った．

(1) ランキング精度向上
WWW 上には，表紙／目次的機能を持つ非常に語数の少ないページと，報告書のような非常に語数の多いページとが混在しており，ランキングの精度に悪影響を与える原因となる．これを考慮し，関連度計算に利用する $tf \cdot idf$ 値を，文書中の異り出現語数を利用して正規化している．
さらに位置索引を利用し，検索条件に含まれる複数の語が文書中の近い位置に出現した場合，その近さの度合に応じて関連度に重みを加える計算式を実装し，ランキングの精度向上を図っている．

(2) 要約文自動生成
CGI スクリプトの機能として，検索結果中にページの要約文を表示する機能を実装した．ここでいう要約文生成は，厳密には抜粋に近い．HTML 文書の特性を考慮して，文書の先頭部分数十～百文字程度を表示すると同時に，当該検索語が出現した部分の周辺文を抜粋して表示する．また，検索語自身は強調表示する．単純な手法ではあるが高速な処理が可能であり，かつその検索語がどのような文脈で利用されているか，が一目で判断できるため，検索結果の内容理解を助ける有効な手掛りとなる．

(3) 関連キーワード機能
WWW 検索においては，ユーザが複雑な論理式を構成することは稀である．代表的な日本語 WWW 検索サイトでも，検索要求の約 9 割が単一語による検索であるという報告もある[4]．これは，Web の普及によって検索に不慣れな一般のユーザが検索システムを利用する機会が増えたことが一因と考えられる．検索条件が単純であるためランキング精度向上には限界があり，ユーザの再検索を支援する機能が重要となってくる．本システムでは，再検索支援機能として関連キーワード提示を実装している．

4 システムの評価

本システムは，(財) 地方自治情報センターの「地域発見」(英語版「Explorer Japan」) をはじめ，当社ホームページの検索システムなど，いくつかのサイトで現在運用中である．特に「地域発見」は地方自治体のホームページを検索対象とするもので，検索対象ページ数は運用中のシステムにおいて最大規模となっている．以下に 1997 年 7 月現在の「地域発見」のデータサイズを示す．

サイト数	670
総ページ数	154,044
うち日本語ページ数	138,236
うち英語ページ数	15,808
テキストサイズ (日)	約 250MB
テキストサイズ (英)	約 30MB

上記データサイズに対して，SparcStation20 相当のマシン上で平均検索処理速度 (検索実行，関連キーワード取得，要約文作成まで) が約 0.6 秒であり，実用上問題ないレベルである．また，システムの集中負荷テストでも，同程度のマシンで同時 40 検索要求まで滞りなく処理できることを確認している．

次に，本システムの特長の一つである関連キーワード機能について，「地域発見」の検索ログから得られた結果をもとに，その有効性を評価する．ログは，1996 年 6 月～1997 年 2 月の 9ヶ月分の日本語検索要求 (95,135 件) を対象とした．図 3にその解析結果を示す．

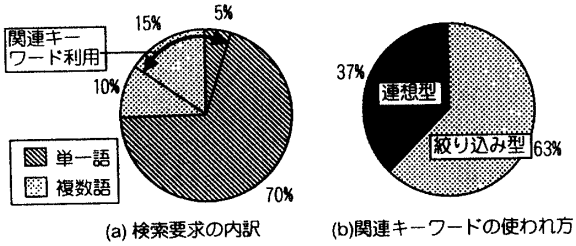

図 3 検索ログの解析結果

これを見ると，単一語による検索は全体の 3/4 程度と依然として多いが，複数の検索語を利用した検索要求のうちの約 6 割が関連キーワードを利用したものとなっており，関連キーワード提示がユーザの再検索支援としてかなり有効に機能していると思われる．

また，関連キーワードの使われ方として 2 種類の方向性があることがわかった．すなわち，もとの検索式のうち 1 語以上と，関連キーワードのうちの 1 語以上とを組み合わせて新しい検索要求とする場合 (絞り込み型) と，もとの検索式を一旦クリアして，関連キーワードのうちの 1 語以上を利用して検索する場合 (連想型) である．比率としては絞り込み型が多いが，連想型も少なからずあり，関連キーワードがユーザの発想支援にも役立っているといえる．

5 おわりに

知的検索ソフトウェア *MEISTER* の WWW 検索への応用について述べた．Web ページの検索においては，ランキングの精度もさることながら，関連キーワードによる絞り込みや連想検索といったユーザ支援機能が重要であり，実際に有効に機能していることが確認できた．

本システムは現状，各 Web ページを一文書としてとらえており，HTML のハイパーテキスト構造は利用していない．しかしながら，アンカー文と URL の参照関係や，URL 同士のリンク関係などを考慮することにより，ランキングや関連キーワード抽出の精度を向上できる可能性がある．今後，この点を中心に改良を行っていく予定である．

最後に，本論文における検索例や評価対象として (財) 地方自治情報センター様が運用する検索ページ「地域発見」を利用させていただいたことを感謝します．

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] 野口喜洋 他: 検索型ナビゲーションを実現したホームページ知的検索システムの開発, 「利用者指向の情報システム」シンポジウム論文集,pp.91-98(1996)

[3] Koster, M.: *Robots in the Web: threat or treat?*, ConneXions, Vol.9, No.4, (1995).

[4] 原田昌紀, 清水奨: WWW 検索システムにおける不特定多数の操作履歴の活用, 情処研報 97-DPS-81-11 (1997).