関連文書検索技術を応用した専門分野ポータルサイトのディレクトリ構築支援

2V－4

永井 明人　増塩 智宏　高山 泰博　鈴木 克志

三菱電機株式会社　情報技術総合研究所

1. はじめに

EC 市場拡大に伴い、宣伝や顧客誘導を狙いとした専門分野ポータルサイトの重要性が高まっている。専門分野ポータルサイトでは、自社ビジネスに関連するWeb ページを広く収集して分類・整理したディレクトリを公開しており、このディレクトリの構築作業は多大な人手に頼っているのが現状である。本稿では、ディレクトリ構築作業で大きなウエイトを占める類似文書の収集に、ベクトル空間モデルに基づく関連文書検索技術を適用し、ディレクトリ構築支援での要件を満たすために施した改良開発について報告する。

2. ディレクトリ構築支援システムの概要

ディレクトリ構築業務における Web ページ収集作業では、ページ内容の適否、特定の言語表現や重要な情報 (掲示板、メールアドレス、リンク集)の有無などの種々の判断基準で、有用な Web ページを収集する。現状では、不要なページを大量に含む全文検索結果に対し、人海戦術により全て目視チェックしているため、非常にコストがかかっている。この作業を効率化するためには、収集目的に合致した類似ページの抽出と、上記のような種々の判断基準による自由な絞り込みが必要である。しかし、従来のベクトル空間モデルに基づく関連文書検索[1]では、特定の言語表現や構造から得られる情報を用いて更に自由に絞り込むことができなかった。

そこで、特定形式情報抽出と関連文書検索を組合わせたディレクトリ構築支援システム(図 1)を開発した。特徴は、(1) 全文検索結果に対して関連文書検索を行なうことで、検索意図に合致した類似ページが効率的に得られること、(2) 特定形式の情報を自動抽出し、更に検索結果の自由な絞り込みに活用できること、(3) 複数文書の適否判断結果や、検索ベクトルの編集(単語の追加・削除、重み調整)結果を用いて関連性フィードバックが可能なこと、などである。

"Portal building support using relevant document search technology"
NAGAI Akito, MASUSHIO Tomohiro, TAKAYAMA Yasuhiro, SUZUKI Katsushi
Information Technology R&D Center
Mitsubishi Electric Corporation

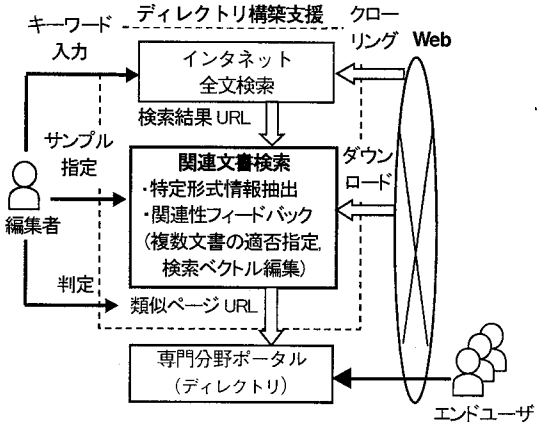

図 1：ディレクトリ構築支援システムの構成

本システムの処理フロー概略は以下のようになる。

(1) 全文検索を用いて Web ページの候補を検索結果 URL として入手する。ここでは、収集対象のページをできるだけ多く集めるために、一般的な単語をキーワードに用いる。
(2) 検索結果 URL から HTML テキストをダウンロードし、関連文書検索に登録する。同時に、掲示板、メールアドレス、リンク集などを情報抽出し、得られた情報を RDB へ登録する。
(3) 自由な文章で関連文書検索を行ない、得られた検索結果に対して、収集したいページのサンプルを指定する。必要に応じて検索ベクトルの編集や、情報抽出された項目により絞り込みを行なう。
(4) 最終的に選択された類似ページ URL をファイル保存し、ディレクトリ構築に活用する。

3. 関連文書検索技術の適用

3.1. 業務要件

本システム開発では、文献[2][3]で用いられているベクトル空間モデル($\chi 2$ 統計方式)に基づく関連文書検索技術を、類似ページの検索に適用した。適用にあたっては、表 1に示すディレクトリ構築の業務要件を満たすように、関連文書検索の改良開発を実施した。

表 1 ：ディレクトリ構築の業務要件

複数ユーザ/複数文書スキーマ化	複数のユーザが、個別の担当分野の文書スキーマで同時に作業できること。
登録文書規模の拡大	全文検索で得られるWebページの候補数を考慮し、一文書スキーマの登録文書数を数万文書程度に拡大する。
処理高速化	夜間バッチで索引登録を行なうことを想定し、索引登録処理は一晩程度で終了するものとする。検索処理は数秒程度の応答時間とする。
リソース占有量圧縮	上記の条件でディスク・メモリ使用量を実用域まで抑制する。

3.2. 改良内容

- 複数ユーザ、複数文書スキーマ対応：サーバークライアント構成をベースとし、クライアントから複数同時に送られる検索・索引登録要求に対して、排他制御/同時実行の管理機能を実装した。
- 登録文書規模の拡大：下記の処理高速化とリソース占有量圧縮を実施して数万文書規模へ拡大した。
- 処理高速化：処理時間の分析を行ない、高負荷の処理に関して以下の高速化改良を施した。
  (1) ディスクアクセス数の抑制
  (2) RDB に対する SQL 実行処理部の高速化
  (3) 形態素解析部の高速化
- リソース占有量圧縮：ディスク使用量に関しては、索引(単語−文書の重み行列)中で類似度計算に寄与しない重みを削除し、索引の構造と検索アルゴリズムを変更した。これによりメモリ使用量の抑制も図った。

4. 評価実験

処理高速化、及びリソース占有量圧縮の効果を確認するために表 2に示した実験条件で評価実験を行なった。実験結果の要点を表 3に示す。表中の改良結果は、改良前の性能との比較である。また、表中の処理速度に関しては、情報抽出された情報をRDBへ登録する処理のオーバヘッド分が含まれている。

表 2：実験条件

マシン環境	PC/AT互換機(サーバタイプ) 2 CPU (Pentium III 850 MHz) メモリ： RAM 2 GB、仮想メモリ 4 GB ディスク回転数： 10000 rpm
OS	Windows NT Server 4.0
RDB	Oracle 8i Workgroup Server V8.1.6

表 3：実験結果

	改良結果
処理速度	約 18 倍に向上
メモリ使用量	約 1/3 に低減
ディスク使用量	約 1/20 に低減

上記の評価結果より、業務要件を満たす性能が達成されていることを確認できた。今後さらに性能改善を行なう場合の課題としては、以下があげられる。

- 処理速度：索引登録処理に関しては、ディスクアクセス数をなくすことが理想であり、全処理を主記憶上で行なうことで高速化が見込まれる。検索処理では、与えられた検索ベクトルに対し、索引の重み行列の中から、類似度を最大化するような文書を求める探索方式の応用が考えられる。
- リソース占有量：索引の重み($\chi 2$ 統計値)の内訳では、微少な重みが大半を占めており、この範囲の重みが、上位の検索結果の精度に与える影響が少ないならば、近似解を得ることを前提にして削除することにより索引サイズの圧縮が可能である。

5. おわりに

ディレクトリ構築業務を効率化するために、関連文書検索技術を適用した支援システムを開発した。実用化改良の評価の結果、業務要件を満たす性能を確認できた。今後は、全体の業務フローに対する効率改善の評価を進めていく。

[参考文献]

[1] 野村, “Concept Base の言語処理と新しいソリューション,” 電子情報通信学会 自然言語処理(NL)研究報告, No. 129, pp. 1-8, 1999.
[2] 藤井, 他 “段落内共起情報を利用した文書自動分類方式,”, 情報処理学会論文誌, Vol. 42, No.3, 2001.
[3] 永井, 他 “CRM における顧客メール分析手法の検討,” 情報処理学会 第 62 回(平成 12 年後期)全国大会 3-81, 2000.