6N－4

ベクトル空間法を利用した WWW上の情報提供サービスの提案*

藤田直毅 西村健士 島津秀雄†

NEC C&C メディア研究所‡

1. はじめに

商品の販売やサービスの提供を行う企業にとっては顧客獲得や顧客サポートのための情報提供サービスが顧客満足度や企業イメージを決める重要なポイントとして位置付けられ,多くの企業でWWWサーバを通じた情報発信を積極的に行うようになってきている.しかしながら,企業のWWWサーバの現状を見ると,十分なサービスを提供していると思われる企業はあまり多くない.

本稿では,必要とされる検索スキルとサービスクオリティに関する問題と,WWW上の情報提供サービスの情報アクセス手段であるブラウジングと検索の問題について考察し,本稿で提案する類似文書検索システムの導入によりそれらの問題が解決することを述べる.

2. WWW上の情報提供サービス

WWW上のHTML文書を対象とした情報アクセス方法としては,任意のキーワードをANDやORなどの論理演算子で組み合わせたブール検索式により全文検索を行なう方法と,内容別に整理された階層的な目次をたどることによりブラウジングを行なう方法が一般的であるが,以下に示すように各々問題点がある.

全文検索手法の問題点

利用者の検索スキルの優劣によって得られる情報が変わってしまうという問題がある他,検索条件の作成が難しい場合や探したいものがはっきりしない場合に使えないという問題がある.その他,コストをかけてもサービスクオリティが向上しにくいという問題がある.

目次ブラウジング手法の問題点

階層的な目次を人手で管理する必要があり,メンテナンスコストがかかるという問題がある.また,大量のデータや高頻度の更新がある場合に迅速に対応できなかったり,目次が適切でなくなるという問題がある.

これらの問題に対処する方法の一つとして,ベクトル空間法を導入するという方法がある.これは,Yahoo (http://www.yahoo.com/) などの全文検索と目次ブラウジングの組合わせをさらに向上させる方法として有力である.ベクトル空間法とは,文書内の単語の出現頻度などを基に文書をあるベクトル空間の一点に対応させ,対応する点と点との距離に基づいて文書の検索を行なう手法である.ベクトル空間法の特徴としては,表面的ではあるが意味的な検索ができること,検索対象文書や質問文の中に構文解析不能な文や図表を含んでいても動作する頑健性を備えていること,適合フィードバックと呼ばれる検索式の対話的な精錬ができること等がある.

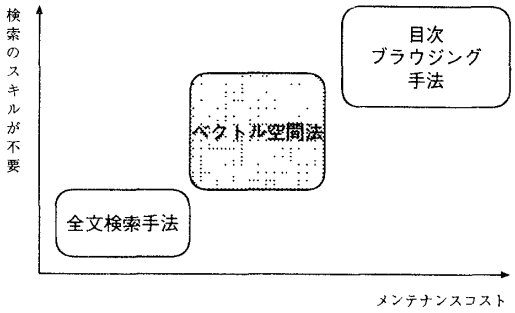

図1: ベクトル空間法の位置付け

本手法は,図1に示すように,必要とされる利用者の検索スキルやメンテナンスコストの視点から全文検索とブラウジングの中間に位置し,全文検索やブラウジング手法の問題点として挙げたものの全てに対してひとつの解答を与えていると考えることができる.現在,ベクトル空間法を利用した検索サービスはExcite (http://www.excite.com/) から提供されている.

3. ベクトル空間法の利用の形態

ベクトル空間法の利用の形態としては,従来の全文検索と似た検索結果のランキング表示の他に,適合フィードバックと類似リンクの2つの形態がある.

適合フィードバックと呼ばれる検索条件の精錬手法

*Information Service Based on Vector Space Model

†Naotake Fujita, Kenshi Nishimura, Hideo Shimazu

‡C&C Media Research Laboratories, NEC Corporation

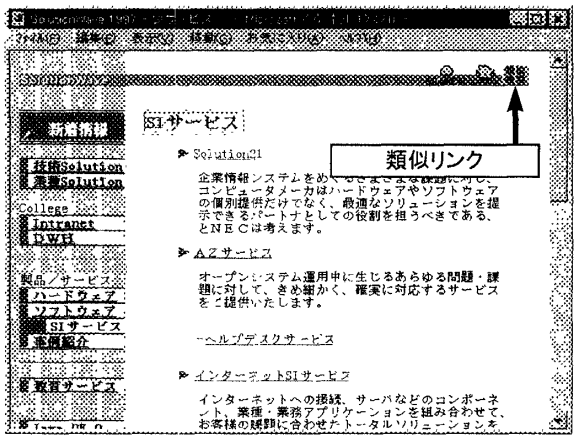

図 2: 類似文書検索システムの画面例 (1)

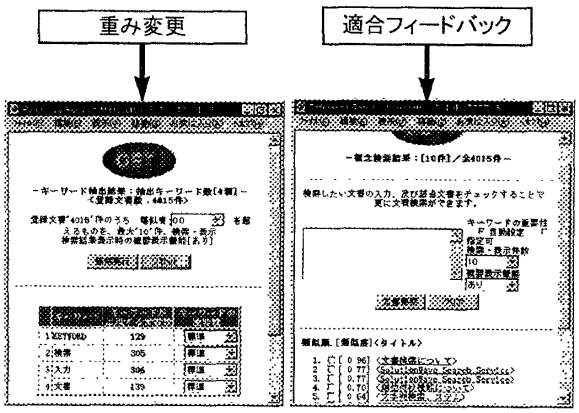

図 3: 類似文書検索システムの画面例 (2)

は, 全文検索で困難だった適切な検索式の作成を容易化し, 検索のスキルを不要とする.

類似リンクは, 指定された文書に類似した文書を検索する機能を持ったものである. 例えばExciteでは[More Like This]という名前のリンクを提供している. これは検索結果の文書タイトルの隣にあるリンクで, これをたどることでブラウジングが可能となっている.

ベクトル空間法を利用したリンク付けの例としては, ハイパーテキストを対象にした研究 [1], USENETニュースグループを対象にした研究 [2] が知られている.

4. 類似文書検索システムの実現例

企業による情報提供サービスでは検索対象文書の数や内容に制限を加えることが可能なため, ベクトル空間法を利用するには望ましい環境である. そこで我々はベクトル空間法を利用した類似検索システムを開発し, 現在 http://www.sw.nec.co.jp/ にて従来の全文検索に追加する形で情報提供サービスを行なっている. 図2の右上部に示すような文書内に埋め込まれた類似リンクボタンによるExciteより直接的なブラウジングが可能なこと, 図3に示すように適合フィードバック時にキーワードの重みを変更できること等が特徴である.

インデックス作成の際は, 予め人手で生成した不要語と同義語を含んだキーワード辞書を作成する. 次に, ロボットと呼ばれるプログラムにより定期的に文書を取得し, 取得された文書から不要なHTMLタグを削除する. さらに, キーワード辞書を利用してキーワードを抽出し, $tf \times idf$ [1] の計算式により, キーワードの重みを予め計算しておく. サービス利用の際は, 入力されたクエリおよび適合文書からキーワードを抽出し, 同様の計算式でクエリベクトルを生成する. 次にクエリベクトルと文書ベクトル間の類似度をコサイン測度を使って計算し, 類似度の高い順に表示する.

利用者は, 図2および図3に示したように, 類似リンクによるブラウジング, 適合フィードバック, 質問文の変更, キーワードの重みの変更を対話的に適宜繰り返すことによって, 必要な文書にアクセスすることができる.

本サービスは現在約4000件の文書を検索対象として, 約4000語のキーワード辞書を利用している. クエリを入力してから結果表示までに5秒～15秒を要する.

5. おわりに

企業のWWW上の情報サービスのためのベクトル空間法を利用した類似文書検索システムの概要を紹介した. 今後は, 全文検索およびディレクトリサービスとの統合について検討を行なう予定である.

参考文献

[1] G. Salton, C. Buckley, On the Automatic Genaration of Content Links in Hypertext. TR 89-993, Cornell University, Ithaca, NY, April 1989.

[2] S. Weiss, S. Kasif, Winnow vs. SMART: A Comparison in the Newsgroup Classification Domain. http://www.cs.jhu.edu/~weiss/ir.html, April 1997.