4K－7

ベクトル空間法に基づく類似文書検索システムの一実現例*

藤田直毅 西村健士 島津秀雄†

NEC 情報メディア研究所‡

1. はじめに

ベクトル空間法は，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，1960年代から実験されてきたSMART[1]が本手法の適用例として有名である．SMARTの応用例は，USENET上のFAQを対象とした検索システムFAQ Finder [2]が知られている．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，対象文書中に構文解析ができないような文や図表を含んでいても動作する頑強性を備えていること等が挙げられる．

本稿では，本手法をヘルプデスクにおける支援システム，特に電子メールを利用して顧客対応を行うシステムへ適用した事例を紹介する．はじめに，ヘルプデスクで日々生成される文書，つまり，問い合わせとその回答との組のQ＆A事例を対象とする検索の問題点について整理し，次に，実際に本手法を適用して開発した類似文書検索システムを紹介する．

2. ヘルプデスクにおける文書検索

現在，ヘルプデスク向けのQ＆A事例を対象とした検索方法としては，対話的な分類属性指定により絞り込みを行なう方法と，任意のキーワードをANDやORなど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が一般的である．

対話的属性指定手法の問題点

本手法を利用するためには，あらかじめ事例の属性として何を選ぶかを決め，各事例について全ての属性の属性値を正確に決める必要がある他，分類されたデータをさらに知識ベースとして構築する必要がある．そのため，作成およびメンテナンスにはエキスパートの参加が必要となり，必然的にデータの更新速度が遅くなるという問題がある．

また，電話を利用したヘルプデスク業務では対話的な処理が有効であるが，電子メールを利用する場合には，かえって操作を増やすという問題がある．

全文検索手法の問題点

本手法を利用するためには，オペレータが検索のスキルを身に付ける必要がある他，オペレータのスキルの優劣によって回答の品質が変わってしまうという問題がある．特に，最新データを提供するために大量の更新データを重複を考慮することなく検索対象とした場合，類似した回答が大量に得られることになり，熟練したオペレータであっても絞り込みの検索条件の作成が困難になるという問題がある．

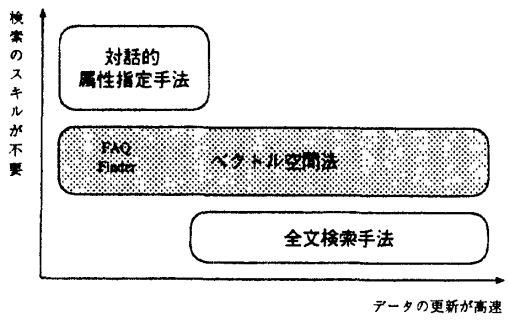


図 1: ベクトル空間法の位置付け

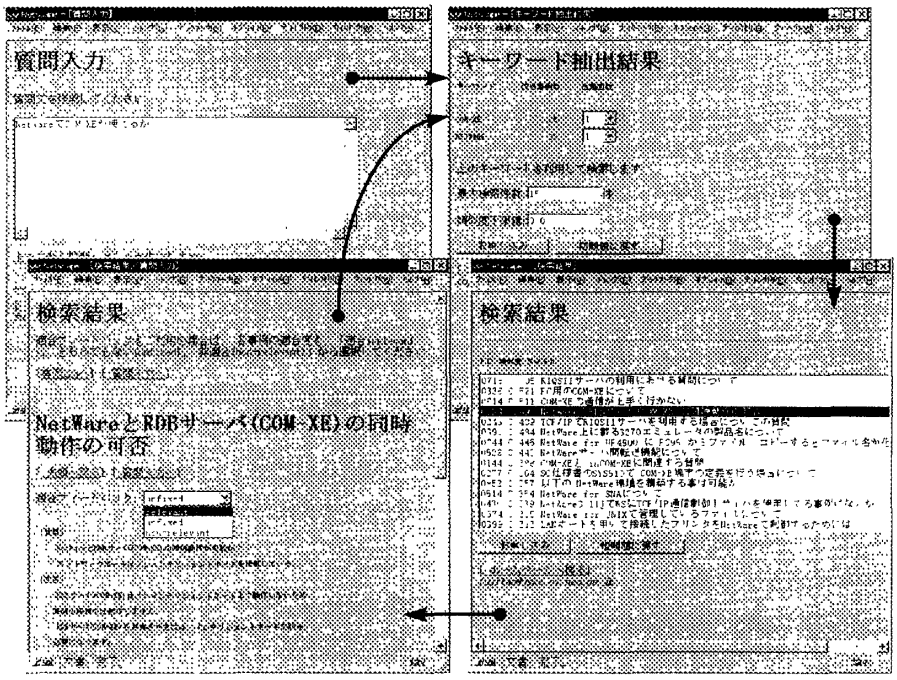
3. ベクトル空間法の適用

前節で指摘した問題を解消するため，ベクトル空間法のヘルプデスク支援システムへの導入を提案する．このことにより，データ更新速度をフリーワード検索の場合と同等に保ったまま，知識ベースを構築せずに，効率的な類似文書検索が可能になる．特に，適合フィードバックと呼ばれる検索条件の精錬手法を備えることで検索条件の作成を容易化し，スキルが不要で効率的な絞り込みを実現する．この他，絞り込みを困難にする類似Q＆A事例を予め機械的に取り除くことが可能となる．

*A Japanese Document Retrieval System Based on Vector Space Model

†Naotake Fujita, Kenshi Nishimura, Hideo Shimazu

‡Information Technology Research Laboratories, NEC Corporation



4. 類似文書検索システムの実現例

本システムは, 事例データベース, キーワード辞書, 事例ベクトル生成部, クエリ入力部, クエリベクトル生成部, 類似度計算部とからなる.

キーワード辞書は事例データベースからキーワード抽出して予め人手で生成する. これには同義語辞書も含まれる. 同時に, キーワードがいくつの事例に出現するかを記したキーワード出現頻度テーブルを作成する. 事例ベクトル生成部は, キーワード辞書を利用して事例データベースの事例からキーワードを抽出し, 事例中のキーワードの出現頻度とキーワード出現頻度テーブルからベクトルの各要素となるキーワードの重み $T\log(N/D)$ を予め計算しておく. ただし, N は事例の総数, T はキーワードの出現頻度, D はキーワードの出現する事例の頻度である.

検索の際は, クエリ入力部により入力されたクエリからキーワード辞書によりキーワードを抽出し, 事例ベクトルと同様の計算式でクエリベクトルを生成する. 次に類似度計算部により, クエリベクトルと事例ベクトル間の類似度をコサイン測度 $\sum_I p_i q_i/(\sum_I p_i^2 \sum_I q_i^2)$ を使って計算し, 事例を類似度の高い順にソートし上位の事例を表示する. ただし, i はキーワードのインデックス, I は不要語でないキーワードの集合, p_i は事例ベクトルの i 番目のキーワードに対応する要素, q_i は検索質問ベクトルの i 番目のキーワードに対応する要素である.

利用者は上図に示す通り, 適合フィードバック (左下画面) およびクエリ変更 (左上画面) や不要語および重みの変更 (右上画面) を適宜繰り返すことによって, 必要な事例を検索することができる.

プロトタイプは現在 905 件のQ＆A事例を検索対象として, そこから抽出された約 3000 語の単語辞書を利用している. WWW 上から CGI 経由で利用した場合, 類似事例のタイトル表示までに 1 秒〜 2 秒を要する.

5. おわりに

ヘルプデスクに蓄積されたQ＆A事例を対象とした類似文書検索システムの概要を紹介した. 今後は, ヘルプデスクでの使い勝手の向上を目指し, 各検索手法の統合について検討を行う予定である.

参考文献

[1] G. Salton, M. McGill, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[2] Hammond, K.J.; Burke, R., and Schmitt, K. A Case-Based Approach to Knowledge Navigation. AAAI Workshop on Knowledge Discovery in Databases. AAAI. August 1994. Seatle WA.