2P−04

情報検索システム IVMap

小中 裕喜　　津高 新一郎　　有田 英一

三菱電機株式会社　産業システム研究所

1. はじめに

PCやネットワークの普及に伴い、電子化文書の流通・蓄積が加速化されつつある現在、大量の文書データからいかにして有用な情報を妥当な時間内に抽出するかが大きな問題となっている。しかしながらキーワード検索や全文検索など、検索式をベースとした従来の情報検索システムでは、検索対象に対して適切な検索式を構成するための十分な知識、経験がなければ、効率的な情報検索は困難であった。

本稿では、そのような問題に対する1つのアプローチとして研究開発を行ってきた情報検索システムIVMap（アイ・ブイ・マップ）[1]について述べる。

2. 情報検索システム IVMap

複雑な検索式を入力することなく、一般的なキーワードからでも効率的な情報検索を可能とするためには、そのような簡単な検索式が往々にして生み出す大量の検索結果に対して、その概要や分布を提示するとともに、利用者による検索の絞込みを容易にする対話的インタフェースを提供することが重要である。

そのためにIVMapシステムでは全文検索などの一次検索機能に加え、i) 各文書を特徴付けるための特徴的キーワード抽出、ii) 特徴的キーワードを利用して一次検索結果を二次元平面に分類配置するマップ生成、iii) マップ上での対話的検索操作を実現するマップ表示・検索インタフェース、の各機能を提供する。以下それぞれの機能について説明する。

2.1. 特徴的キーワード抽出

文書タイプごとに定まるフィルタにより切り出された、キーワードを抽出すべきテキスト領域から、検索対象となる文書集合の中で、各文書の特徴を表すと思われるキーワードをオフラインで抽出する。

キーワード抽出には形態素解析を用いていない。まず漢字やカタカナといった字種に関する情報をもとにキーワード構成要素を抽出する。また字種では効率よく抽出できないものについては、別途登録された情報にマッチするものをキーワード構成要素として抽出する。そしてテキスト上連続したキーワード構成要素を連結してそれぞれキーワード候補とした上で、各候補について不要な文字列の処理を行う。

不要文字列処理は、各キーワード候補から各文書の特徴を表すのに不要な文字列を削除することを目的とした重要な処理であり、各候補の文字列全体、先頭、末尾に関して予め登録された必要文字列、不要文字列の情報を利用し、必要文字列にマッチせず不要文字列にマッチする文字列の削除を繰り返す。形態素解析を介さないため、例えば不要な末尾文字列に関しては、一般的な接尾語（例えば技術文書における「...手段」「...機能」など）の他に、漢字で始まる副詞が名詞に連なる場合（例えば「...の分散管理特に...」）なども考慮する。残った各候補の文字列には、類義語情報に基づく置換が行われ、最終的なキーワードとなる。

上記処理に用いられる情報はすべて正規表現によって表され、柔軟性の高い記述が可能となっている。またこれらの情報は一般に検索対象となる文書集合の属する分野や文書タイプに応じて異なる。そこでこれらをモジュール化しておき、例えば計算機関連特許を処理する場合には、一般技術文書用モジュールに計算機関連文書用モジュールと特許用モジュールを組み合わせて用いるといった、モジュールの統合を可能にし、それぞれの情報の再利用性を向上させている。

2.2. マップ生成

各文書から抽出された特徴的キーワードの集合によって構成されるベクトル空間上の文書間の類似度に基づき、類似した文書が同一／近傍の場所に配置されるよう、一次検索結果を六角形のセルからなる二次元平面にマップ化する（トポロジカルマッピング）。

マップ生成にはBatch Map 型SOM [2]に類した高速アルゴリズムを用いており、例えば10000件の特許文書でもマップ生成自体は最新のPCで数秒程度で完了し、対話的な情報検索が可能となっている。

2.3. マップ表示・検索インタフェース

二次元平面のマップに分類配置された文書集合の平面図／鳥瞰図を、各セルに配置された文書群を代表するキーワードとともに表示し、利用者が検索を進めていくための各種操作インタフェースを提供する。

デジタル通信関連特許に対して“コンピュータ”

The IVMap information retrieval system
Hiroki KONAKA, Shin'ichiro TSUDAKA, Hidekazu ARITA
Industrial Electronics & Systems Laboratory
Mitsubishi Electric Corp.

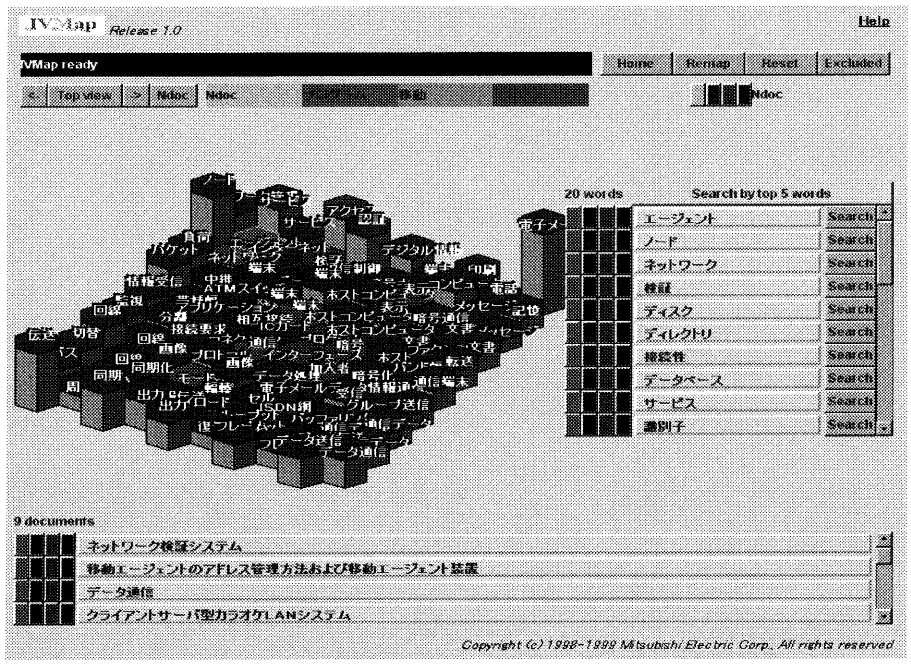

図 1　IVMap 画面例

で全文検索を行った結果約 2500 件の文書を 10x10 のマップに分類配置した場合の鳥瞰図を図 1 に示す。鳥瞰図においては、各セルに表示される六角柱の高さがそこに分類配置された文書の数を相対的に表している。

各セルをクリックすると、そこに分類配置された文書群の特徴的キーワードが上位 20 個まで画面右に表示され（最上位はマップにも表示）、文書群の概要の把握をサポートする。また画面下部には各文書のタイトル一覧が表示され、適当なタイトルをクリックすることにより内容を閲覧することが可能となっている。

画面右の各キーワードの左側にある三原色のボタンを押すと各キーワードに色が割り当てられ、色のついたキーワードを含む文書が配置されたセルにもその色がつけられる。このとき、例えば赤のキーワードと緑のキーワード双方に関連したセルは黄色く表示されるというように、色づけされたキーワードに関連したセルに関する AND/OR 検索が、色の重ね合わせにより視覚的に実現されている。

また新たに興味をもったキーワードの右側の **Search** ボタンを押したり、一番上の **Search by top 5 words** ボタンを押すことにより、それらのキーワードを用いて一次検索をやり直すことも可能となっている。

興味深そうなセルをダブルクリックでいくつかマークし、画面右上の **Remap** ボタンを押すことにより、マークされたセルに配置された文書からなる新たなマップを生成することも可能である。これにより、例えば多くの文書が配置されたセルをさらに掘り下げてマップ化し、Scatter/Gather [3] のように階層的に検索を進めていくことも可能となっている。

3. おわりに

IVMap システムはクライアントサーバ型で実装されており、ネットワーク経由で文書サーバを共有しながら WWW ブラウザより検索を行うことが可能となっている。

IVMap システム自体は網羅的な検索を直接サポートするものではなく、むしろ利用者の検索要求が曖昧な段階においても、一般的なキーワードから検索を進めることが可能なシステムを目指している。利用者は必要に応じてマップを生成しながら、一次検索結果の概要や傾向、分布を把握するとともに、興味深いキーワードの表示されたセルやその近傍のセルを探索し、興味深い文書を発見していくことが可能である。また新たに興味をもったキーワードを用いて検索の方向性を修正したり、興味深い一部のセルに限定してより詳細なマップを作成し、さらに検索を絞り込んでいくなど、柔軟で発見的な情報検索が可能となっている。

参考文献

[1] 有田, 安井, 津高, *単語集合の自動構造化機能を持つ「情報散策」方式*, 情報処理学会 自然言語処理研究会, 108-11, pp.69-74, 1995.

[2] Kohonen,T., *Self-Organizing Maps*, Springer-Verlag, 1995.

[3] Cutting,D.R., Karger,D.R., Pedersen,J.O., Tukey,J.W., *Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections*, Proc. of ACM SIGIR'92, pp.318-329, 1992.