情報検索システム「Datahunter」におけるビジュアライズ機能

3P－11　坂倉 健太郎　吉川 耕平　西村 英樹　稗田 薫

シャープ株式会社 技術本部 マルチメディア推進本部 システム開発センター

1. はじめに

パーソナルコンピュータの普及やネットワークの発展に伴ない、ユーザがアクセス可能な電子文書の量は爆発的に増加している。そこで、大量の文書群の中から、ユーザの所望する文書を効率的に高速に検索する研究が行なわれている [1]。

一般的な文書の検索手法として、ある単語を指定しその単語を含む文書を探し出すものがある。しかしこの方法では、検索に用いる単語を適切に選ばないと期待する結果は得られない。一方、単語ではなくある文書を指定して、それと類似する文書を検索するという手法がある。この場合、ユーザは適切な単語を考える必要がない。

我々は、文書で文書を検索し、その結果を様々な形でビジュアライズし提示する情報検索システム「Datahunter」を開発した[2]。本論文では従来の検索システムの問題点を、検索結果の提示という観点から分析し、これを解決するために Datahunter で取った手法を紹介する。

2. 文書検索と検索結果の提示手法

多くの検索システムは、検索終了後、単に結果をリストで提示する。文書による検索の場合は、キーとなる文書と検索対象文書に含まれる単語の出現頻度から個々の文書がキー文書とどれだけ近いかを示す文書間の類似度を計算するものも存在する。しかしいずれの場合もユーザは、結果を一つずつ内容確認し取捨選択せねばならない。この方法では、結果の内容確認という作業を何度も繰り返した後で初めて検索キーが不適切であることを知ったり、結果リストの量によっては内容確認自体が困難になることが起こる。また、文書で文書を検索する場合、自分がどういった検索を行なったのかを明確に把握することができないため、適切な検索処理が実行されているかどうか不安を持つことがある。以上のことから、検索結果の文書のリストを単に類似度順に並べるだけでは不十分である。

そこで、検索結果と合せて以下のような情報も視覚的に提示すれば問題を解消できる。

- 検索対象となる母集団全体の時間的な分布
- 各文書における検索キーとの類似度と、作成(あるいは更新)日時との相関性
- ある文書を端的に特徴付ける単語
- 母集団における単語の出現頻度の時間分布

これらの視覚化により、検索結果の文書が母集団の中でどう位置付けられるのか、どのような特徴があるのかを直感的に把握できる。従って単に類似文書を検索するシステムよりも、ユーザは検索結果を納得して受け入れることができる。

次章では、上に挙げた情報をビジュアライズする方法を説明する。

3. Datahunter におけるビジュアライズ機能

システムは、単に文書で文書を検索できるだけでなく、その検索結果全体や個々の文書の特徴を、より容易に理解できるよう様々な形でビジュアライズする機能を持つ(図 1、図 2)。

3.1 モザイクを用いたビジュアライズ

文書による検索では、検索キーの文書と母集団に属する各文書との類似度を計算する。システムは類似度の高い順に文書のタイトルなどをリスト表示する一方で、各文書の類似度と作成時刻の相関を二次元画面上に表示する。これはモザイク表示と呼ば

Visualization of Information Retrieval System "Datahunter",
Kentaro Sakakura, Kohei Yoshikawa, Kaoru Hieda, Hideki Nishimura,
Sharp Corporation, Corporate Research and Development Group

れる検索結果の視覚化方法の一つである。

図1は検索結果表示画面の一例である。図の上半分はモザイク表示の部分である。モザイク表示部分では、横軸は文書の作成時刻を示し、縦軸は類似度を示す。ここにすべての検索対象の文書を、それが作成された時刻と検索キーとの類似度によって該当するマスにプロットする。各マスに該当する文書の量に応じてマスの色を変えることで、どの程度内容が類似した文書が、どの時刻にどれだけ作成されているかを一目で知ることができる。

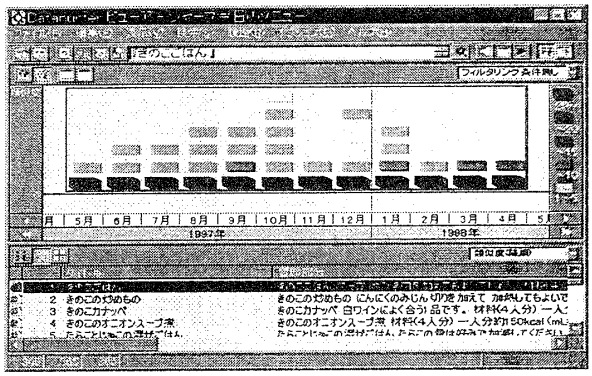

図1：モザイク表示画面

次に、モザイク表示によって得られる効果を説明する。例えばモザイク表示画面の上部(類似度の高い部分)に、文書が多いことを示す色のマスが集まっていれば、内容的に似た文書が非常に多いことが視覚的に即座に把握できる。しかしこの場合でも、モザイク画面上部のマスがほとんど左側に集まっていれば、類似度の高い文書がいくら多くても、大半が何年も前に作成された文書であることがわかり、最近の文書をだけを閲覧すればよいと判断できる。このように、単に検索結果をリスト表示する従来の手法とは異なり、検索と同時に視覚的な判断材料が与えられるため、逐一リストの内容を確認する必要がない。

画面の下半分は検索結果を類似度順に並べたしたリストである。リスト中の各項目を選択すると、その文書がプロットされたマスの色が明るく変化する。つまり、選択した文書が母集団の中でどの位置に属しているか、類似度と作成時刻の二つの側面から簡単に知ることができる。

3.2 その他のビジュアライズ機能

Datahunterにはモザイク表示の他にも文書中に含まれる単語についてその文書を端的に特徴付ける度合を提示するキーワードビュー機能や、ある単語の母集団における分布を折線グラフで表示する機能(図 2)があり、より説得力のある検索結果をユーザに示すことができる。

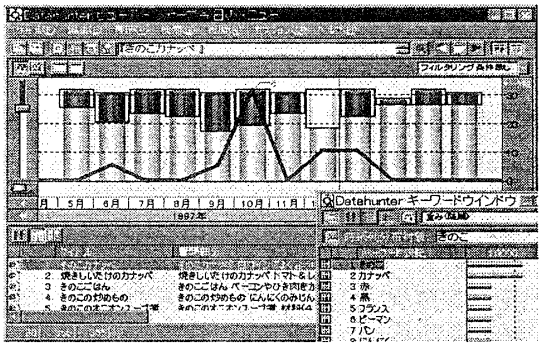

図2：キーワード表示画面

4. おわりに

以上のようなビジュアライズ機能を持つことで、Datahuner の「文書で文書を検索する」という特長機能をより活かすことが可能となる。特にモザイク表示は従来ユーザには見えなかった検索結果の構造を直感的にユーザに示すことで大量の電子文書の中から所望の文書を取り出す作業の効率を上げる機能である。

今後は、電子文書だけでなく、動画像や音声などのディジタル情報をターゲットとした検索システムにおける検索結果の提示方法についても研究していく予定である。

参考文献

[1] 武田英明, “ネットワークからの知的情報収集・統合”, 日本ソフトウエア科学会, Dec. 1996

[2] Kohei Yoshikawa, Heiichi Yamamoto, “Whole Data Visualization by Similarity Degree for Mobile and Adaptive Information Retrieval” International Conference on PDPTA, Jul. 1999