テキスト情報の可視化による情報検索

武田 浩一
日本アイ・ビー・エム株式会社 東京基礎研究所
神奈川県大和市下鶴間 1623-14
takeda@trl.ibm.co.jp

1 まえがき

WWW(World-Wide Web) の急速な普及によって、我々の情報検索に対するイメージは大きく変わりつつある。具体的には、ハイパーリンクをもつ HTML 文書 [4] を中心として、世界中のテキスト情報と膨大な量のマルチメディア・データが WWW ブラウザから日常的に利用可能となった。分野別に整理されたディレクトリ・ガイドから、各種の全文検索エンジンや、サイトを渡って情報を収集するエージェントのようなものまで、多様な情報要求を満たすための技術が盛んに研究・開発されている。

HTML 文書では、画像や音声が多様な情報表現に大きく貢献しているが、テキストによって表現された情報を解析・理解することが依然として非常に重要である。通常このような情報は大部分が非構造化された自然言語の文であり、適切な自然言語処理技術の援助がない限り、日本語のように単語の境界が自明でない言語はもちろん、英語でも検索結果の精度や、その重要度の順序に満足できないことになる。さらに、最近では、情報を提供するサイトが極端に増加しているため、検索結果を絞り込むことや、興味のある内容をそこから見つけだすことは相当手間のかかる作業になる。したがって、テキストの分類 [2]、テキスト情報のフィルタリング [15]、要約 [14]、ダイジェスティング [9] といった自然言語処理や、膨大な内容のテキスト情報を読む代わりに効果的な可視化 [8] を行なうという技術が必要である。

従来の情報検索では、自然言語処理として統計的な性質やシソーラスの利用を含めた検索の妥当性が非常に重視されてきたが、可視化による効果は上記の問題の解決にかなりのインパクトを持つものと考えられ [7]、自然言語処理と可視化の技術をうまく組み合わせることが鍵となる。

このことは、電子図書館 [13, 1]、電子取引や電子出版 [12] という分野でも、電子化された情報を効果的に作成/共有/配布するための極めて本質的な問題といえ、教育から広告、出版までの広い範囲にまたがるものである。次節では、このように情報の検索と、望む情報へ効率的に到達できるようなナビゲーションについて論じ、可視化の役割を 3 節で説明する。4 節では、これらの考えに基づき我々が提案している Information Outlining[7] という手法について述べる。

2 ネット社会における情報の検索

HTML 文書を主要な情報表現形式とした情報検索では、従来の検索式による対象の同定に加えて、ハイパーリンクをたどるというナビゲーションの要素が微妙に組み合わされている。これは、ユーザが検索する手段として、例えばキーワードを指定して興味のあるデータ集合を定義する方法と、リンクをたどって個別のデータに順に調べていく方法が混在することを意味する。このような情報の検索は、**特定検索**と**広域検索**の 2 種類に大別できるものと考えられる。特定検索は、ある事実の記述の有無や、リンクをたどって、ある日時に起こった出来事を調べる、といった要求であり、限定的、個別的であり、非常に詳細な情報を必要とすることが多い。このため、膨大なデータに対しても、全データを検索し、かつフィルタリングや要約による抽象化を避けないと目的とする情報が得られない場合がある。このような検索のコストは非常に大きなものになる可能性があるが、ユーザにとっても、求める情報が検索コストに見合うほど貴重なものであることも多いため、効果的な検索データの絞り込みやナビゲーションが欠かせない。ハイパーメディア・システムでの、いわゆる迷子 (lost in space)[3] の問題も、このようなリンクの個別性によって生じる問題である。

一方、広域検索は、あるデータの集合が示す傾向を調べたり、「情報散策」[10] といった、特に明確な目的のない検索のために大まかなデータの集合を指定するといった情報要求である。特定検索と違い対象となるデータのサイズが大きくなるほど信頼度が増すような市場調査や、意思決定に必要な資料を作成するのにも広域検索が使われる。したがって、広域検索では、膨大なデータをサンプリングしても有効な情報を得られたり、フィルタリングや要約による抽象化も効果的に利用できることが多い。また、可視化は、このような傾向や思いがけない事実の発見を容易にするための強力な手段となる。

現実の情報検索では、特定検索と広域検索の要素が絡み合うこともあるため、上記のような効果的なナビ

ゲーションと可視化を組み合わせて、様々な情報要求に答えられるようなインタフェースを構築することが望まれる。このような研究例には、植田らのCastingNet[11]がある。CastingNetでは、ハイパーリンクを一般化し、データを動的に対応づける関係軸というn項関係が定義されている。関係軸ごとにハイパーチャートと呼ばれる、種々の可視化ツールが提供され、関係軸がもつ情報は、3次元グラフや系統樹といったわかり易い表現方法によって表示される。ユーザはハイパーチャートとデータとの間を双方向に参照することができ、可視化とナビゲーションが巧みに統合されている。

ハイパーリンクをn項関係に一般化したり、ハイパーリンクを検索式に置き換えて動的に評価できるようにすると、従来のハイパーメディア・システムの特定検索的な性質は薄れていく。すなわち、ハイパーリンクとは、検索式と検索式を満足する要素の順序関係を定義するものに一般化でき、検索式に対する検索結果の関連度を扱えるような従来の情報検索システムとの差異がほとんど感じられなくなる。ただしConklinが指摘したハイパーメディア・システムの「認知的なオーバヘッド」[3]の問題は依然として残り、ユーザが自然言語によって、効率的に処理が可能な検索式を表現できない限り、求めるデータの集合をちょうど規定するような都合の良い単語や句のブール式が見つからないケースが多く発生すると考えられる。従来の情報検索では、この問題には、シソーラスなどを利用した上位語の利用や、検索式の拡張(query expansion)と関連度フィードバック(relevance feedback)[5]という手法などが提案されているが、可視化の分野では、より高度なブラウジングと対話処理[8]で解決しようとしている。これらの研究が示唆することは、ハイパーメディア・システムにしても情報検索システムにしても、

- 最初に、興味のあるデータ集合を規定する方法
- 与えられたデータ集合の特徴をわかり易く表示する方法
- 与えられたデータから、それに関連するデータを効率的に参照する方法
- 与えられたデータ集合をより効果的に絞り込む方法

という4つの要素が本質的であるという認識である。可視化が検索式やハイパーリンクの限界を補うのに有望視されるのは、可視化によって把握できる情報が、上記の4つの要素すべてに関わるからである。

3 テキスト情報の可視化

データベースの分野では、最近多次元データベースやデータ・ウェアハウス[6]といった、データの多角的利用や表計算アプリケーションなどとのインタフェースを考慮した新しいシステムの概念が現れた。関係データベースは、データをn属性の表と考えれば、比較的容易にビジネス・グラフィックスや、データ可視化ツールと組み合わせて利用することができる。面白いことに、このようなデータベースとそのアプリケーションとの関係は、テキスト情報とその可視化や情報検索インタフェースにもそのまま当てはまる。これは、テキストが語や文、段落といった要素から階層的に構成されているためで、どの要素、あるいはどのような構造に注目するかで、テキストから得られる情報が変化するためである(図1参照)。この性質を利用してテキスト情報を様々な用途に利用することが可能となる。

自然言語処理	抽出される情報	可視化
形態素解析	単語	単語の出現頻度
	共起関係	関連語・KWIC表示
+シソーラス	概念分類	重要語とその分布
		概念階層
		テキスト分類
統語解析	係り受け	統語的n項関係
	文の重要度	要約
意味解析	語義	重要語の意味分類
	意味関係	意味的n項関係
意味理解	文の意味内容	各種の意味関係
文脈理解	段落の内容	話題の流れ
	接続関係	

図1: 自然言語処理と可視化

テキスト情報が階層的であることは、その可視化の方法も、要素ごとに多数存在することを意味する。テキスト情報を可視化する目的は様々であるが、通常は、膨大な量のテキストをそのまま読むかわりに、明示的ではない次元(属性)の追加、あるいは情報の配置によって、そのテキスト情報に特定の性質や、全体像を理解するための補助情報をユーザにわかり易く提供することである。このような個別の表現手段をビューと呼ぶことにする。CACMの1995年4月号の電子図書館特集では、比較的低レベルの自然言語処理と、文書のカテゴリや検索語との一致度といった情報を可視化に利用した技法が紹介されているが[8]、テキストに含まれる時間的あるいは空間的な関係を用いると、そのテキストの特徴を極めてわかり易く表示できることがある。これらの表示は、テキストの意味内容や、分野に応じて、向き/不向きの差が激しいことも経験的にわかっている。

したがって、テキストから抽出された情報を可視化する方針は、部分的な情報とそれをわかり易く表現するビューを、できるだけ豊富に提供し、これらをユーザ

が自由に参照することで、テキストの分野依存性やユーザの個別な要求に対応できるようにすることと結論づけられる。また、これらのビュー相互の関連も同様にわかり易く表現できれば、ナビゲーションにも役立てることができる。全文検索が常識的になった最近では、形態素解析レベルで得られる情報は、検索を高速化するためのインデクスとして管理されることが多い。統語解析以上の情報は、その精度と物理的な記憶スペースの問題があり、まだ有力な方法が存在しない。長尾が指摘するように[13]、タイトルや各章の表題といった、そのテキストを表現する代表的な情報が得られる場合には、これらの情報を意味のレベルまで解析して、高度なインデクスを作成しておけば、検索機能・時間と記憶スペースのトレードオフへのよい解決法となる。ただし、最近の全文検索エンジンの記憶量、処理速度、およびインデクスの更新頻度を考慮すれば、従来では考えられなかったほどの豊富な情報が、事前にテキストから抽出できると過程してよいかもしれない。

4 Information Outlining

Information Outliningとは、我々が電子図書館のプロトタイプに採用したアイデアで、これまでに述べてきたような自然言語処理と可視化の技術に基づいて、検索結果の種々の概観、絞り込み、ナビゲーションを実現するための、総合的な検索の枠組である。

1. 地理別件数分布や年別件数分布といったビューを多数用意する
2. ビューごとに検索結果をさらに絞り込む個別の機能と、他の関連するビューを参照する機能を定義する
3. 各ビューは、検索結果が変化するのと同期して動的に変化する
4. ビューで表現できる情報に対応して、自然言語処理や情報抽出の技術を利用した計算機構を用意する

ビューを通して見えるデータは、検索結果の部分集合であり、かつ、その部分的な情報が表示されている。したがって、ビューをデータの集合と同一視して、ビュー同士の集合演算や、あるビューから距離kのリンクで到達可能なデータの集合を求めるような推移的閉包演算が定義できる。ユーザはこれらの操作を、ビューや興味のある対象を選択するような視覚的インタフェースを通して、簡単に実行できる。単純な可視化とInformation Outliningが大きく異なるのは、このようにビューを、さらに検索に利用するための能動的なオブジェクト集合として扱う点にある。これにより、検索式だけでは、なかなか表現しにくいようなデータ要求、従来の情報検索では不十分であったナビゲーションを支援すること、分野依存性やユーザの個別性に対応すること、などを可能にしている。

また、Information Outliningでは、シソーラスや人名辞典、企業情報などの分野別のリソースを、特定のビューと対応づけることで、元データだけからは得られないような情報を、巧みに補いながら、より高度な検索や、可視化に自然に利用することができる。たとえば、姉妹都市の情報が図2の日本地図ビューに対応づけられているとすると、特定の話題のデータ集合に現れる、都市別の姉妹都市との友好事業の一覧を抽出・表示することが可能になる。

図2に、我々のプロトタイプの2つのビューを示す。左下の日本地図ビューでは、検索結果で言及されている日本の都道府県名を、その件数別に5段階に色を変えて表示している。このビューの上位ビューとして、世界地図ビューと相互に参照が可能である。右下の月別分布ビューでは、検索結果の各データの日付に基づいて、月別の件数を棒グラフで表示している。下位ビューとして、さらに特定の月の日別分布ビューがある。この月別分布ビューからは、検索件数が、右上がりに増加しており、検索データの時間的な分布の特徴(例えば、話題性が非常に高まってきていること)が直ちに理解できる。この例では、日経NEEDSの1994年度の日経新聞記事約15万件を検索対象とし、「インターネット」という用語を含む記事を検索しているところである。これらの各記事には、あらかじめシソーラスに定義されている分野別キーワードが数十個程度与えられており、この情報を可視化に利用している。ただし、我々のプロトタイプでは、形態素解析技術とシソーラスによって特許申請やその他の一般のテキストからも、有効なキーワードが抽出できることを検証している。

我々のプロトタイプでは、最初に興味のあるデータの指定は従来のキーワード検索あるいは全文検索によるが、ビューによる可視化を利用して、2節の最後に述べたような4つの要素を取り込んでいる。また、全く検索式を用いなくても、サンプリングによって、一定数のデータを選びだし、上記のようなビューを使って検索を進めることもできる。

5 あとがき

本論文では、ネット社会が我々にもたらす情報の海を航海する手段として、自然言語処理と可視化によって強化された情報検索手法を論じた。マルチメディアがこれほど豊富に使われるようになっても、テキスト情報の柔軟かつ高度な可視化が、大量の情報から求めるデータを探し出すための鍵であることに変わりはなく、いわゆる

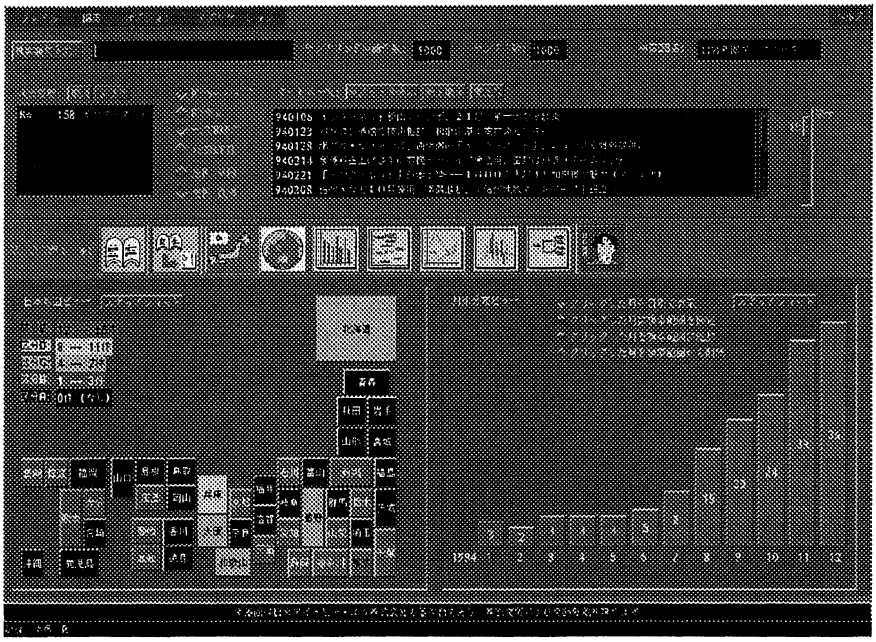

図 2: 日本地図ビュー (左下) と月別分布ビュー (右下)

scientific visualization 中心の可視化の技術が、textual information visualization を含めて発展していくことが今後の大きな流れであろう。

謝辞

日頃御討論いただく東京基礎研究所の研究員諸氏に感謝いたします。

参考文献

[1] ACM. "Communications of the ACM Special Issue on Digital Libraries". *Communications of the ACM*, 38(4), Apr. 1995.

[2] N. J. Belkin and W. B. Croft. "Information Filtering and Information Retrieval: Two Sides of the Same Coin". *CACM*, 35(12):29–38, Dec. 1992.

[3] J. Conklin. "Hypertext: An Introduction and Survey". *IEEE Computer*, 20(9), Sept. 1994.

[4] D. Connolly. "Hypertext Markup Langage (HTML): Working and Background Materials". http://www.w3.org/hypertext/WWW/MarkUp/MarkUp.html.

[5] E. N. Efthimiadis. "A user-oriented evaluation of ranking algorithms for interactive query expansion". In *Proc. of SIGIR'93*, pages 146–159, Pittsburgh, PA., July 1993.

[6] J. Hammer, H. Garcia-Molina, J. Widom, W. Laio, and Y. Zhuge. "The Stanford Data Warehousing Project". In *IEEE Data Engineering Bulletin*, June 1995.

[7] M. Morohashi, K. Takeda, H. Nomiyama, and H. Maruyama. "Information Outlining - Filing the Gap between Visualization and Navigation in Digital Libraries". In *Intl. Symp. on Digital Libraries*, pages 151–158, Tsukuba, Japan, Aug. 1995.

[8] R. Rao, J. O. Pedersen, M. A. Hearst, J. D. Mackinlay, S. K. Card, L. Masiner, P.-K. Halvorsen, and G. G. Robertson. "Rich Interaction in the Digital Library". *Communications of the ACM*, 38(4), April 1995.

[9] 佐藤, 佐藤: "ネットニュースのダイジェスト自動生成". 言語処理学会第 1 回年次大会, pp.297–300, 1995 年 3 月.

[10] 諸橋, 堤, 丸山, 野美山: "情報検索システムにおける効果的なナビゲーション機能の提案". 「ディジタル図書館」ワークショップ論文集, pp.45–49, 1994 年 11 月.

[11] 植田, 増田, 石飛: "CastingNet: 情報の組織化と可視化ブラウジングのためのハイパーメディアシステム". コンピュータソフトウェア, 12(4):56–71, 1995 年 7 月.

[12] 石塚: "ディジタル図書館における基本出版技法: SGML". 「ディジタル図書館」ワークショップ論文集, pp.3–14, 1994 年 11 月.

[13] 長尾: "電子図書館". 岩波書店, 1994 年.

[14] 渡辺: "新聞記事の要約のための一手法". 言語処理学会第 *1* 回年次大会, pp.293–296, 1995 年 3 月.

[15] 藤澤, 絹川: "情報検索における自然言語処理". 情報処理, 34(10):1259–1265, 1993 年 10 月.