ビュー合成によるテキスト情報の可視化

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

情報可視化の目的は，陽には認知できない情報を理解しやすい形で表現することにあり，情報検索や電子図書館のGUIとしてもクラスタ，検索式との関連度，参照構造などの様々な情報の表現に利用されてきた．[6] ただしテキスト本来のもつ情報は，語の分布，ストーリー展開，本の3次元的な可視化といったいくつかの限られた分野を除けば，あまり活用されてこなかったといえる．ところが最近になって誰でも大量のWWWページを検索できたり，社内の膨大な文書データを共有できるようになったため，以下のようなテキストからの情報抽出や情報可視化への極めて強い需要が発生するようになった．

- 情報検索において，関連度フィードバックのような検索結果の改善や絞り込みのためのインタラクションを，要約技術と可視化によって効率化したい
- インターネットを通して得られたユーザのフリーコメントなどをもとにテキスト・マイニング技術 [10] と可視化によって，企画やマーケティングに活用したい
- 携帯端末や車載用情報機器のように，個別の情報要求 (道案内，レストランガイドなど) を満足する情報を，情報フィルタリングと可視化によって制限された画面で表現したい

これらの需要には，Shneidermanの「まず概観 (overview) せよ．そしてズームやフィルターを行い，必要に応じて詳細に注目 (details-on-demand) せよ」[7] という視覚的な情報検索の定型パターンが適合している．このことは，自然言語処理を主体とする情報抽出と情報視覚化の共通のパラダイムによって，上記のようなタスクを含む多様な問題の解決が可能となることを意味する．

本論文では，従来の可視化技術をテキストのもつ様々なレベルの情報に適用し，上記のような問題解決を可能とするような情報可視化のモデルを提案する．テキストがもつレベルの異なる情報に多様な可視化表現を適用する場合には，これらの可視化表現を合成することで，より情報の多い，あるいは理解しやすい可視化表現 (ビュー) が得られることがある．ここで提案する情報可視化のモデルには，このようなビュー合成の機能が含まれており，それが具体的にどのように応用できるかについても述べる．

2 情報可視化のモデル

情報可視化には，データ型とタスクの組合せに応じて各種の技法が知られている．まずテキスト情報を含んだデータ集合を可視化するときに概要，ズーム，フィルターに相当するデータとその操作を以下に列挙する．

自然言語処理	抽出される情報	可視化
形態素解析	単語 共起関係 named entity 統計情報	単語の出現頻度 [5] 関連語・KWIC 表示 [6] 人物，組織名など 自己組織化マップ [3]
＋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 [6]
統語解析	係り受け 文の重要度	統語的 n 項関係 要約 パラグラフの表現 [1]
意味解析	語義 意味関係	重要語の意味分類 意味的 n 項関係
意味理解	文の意味内容	文の意味分類 言い換え
文脈理解	段落の内容 接続関係	話題の流れ

図 1: 自然言語処理と可視化

自然言語処理の複雑さと可視化できる情報の複雑さはある程度の相関があるといえる．局所的かつ統語的な情報は大量かつ高速な処理が可能であり，文書集合全体の可視化に対して強力な手段を提供する．意味的あるいは文脈的な情報は一般に高価な処理によらないと獲得できないため，特定のテキストの内容のような限定的な対象に適用するか，あるいはバッチ処理によってオンライン検索時に使用する情報を事前に計算しておく必要がある．可視化のタスクでいえば，前者は主に概要の可視化，

後者は詳細の可視化によく適合する．[1]概要のレベルでの主要な機能は興味のあるデータ集合の特定や全体の傾向，大まかな特徴などを把握することであり，トップダウン的なナビゲーションの機能やデータ分析の機能が含まれる．詳細レベルでは，個別のテキストの内容をもとに，これに類似した内容をもつ文書の検索，関連情報や参照情報の表示，要約や翻訳といった機能が必要となる．

これらのレベルの異なる情報抽出と可視化を統合するために，次のような層状の可視化のモデルを提案する．

1. 与えられた複数のデータ集合は，1つ以上の可視化のレヤーを線形のキュー (queue) に重ねることによって可視化される．各レヤーには対象とするデータ集合と，それを可視化する表現が対応づけられ，レヤーの重ね合わせによって情報が漸進的に表現されていく．

2. レヤーは，1つの可視化表現に対応する基本レヤー，n個の可視化表現に対応する分割レヤー，データ集合の特定の部分に焦点を当て，選択するためのズーム・レヤーの3種類がある．分割レヤーは，1つのキューをn個のキューに分離する．

3. 2つのレヤー L_1, L_2があり，L_1の上に L_2が重ねられるのは，
   - 両者が合成可能な基本レヤーの場合
   - L_1が基本レヤーで，L_2がその部分集合を選択するズーム・レヤーの場合
   - L_1が基本レヤーで，L_2がそれをn通りに可視化する分割レヤーの場合
   - L_1が分割レヤーで，L_2がその1つの可視化表現に合成可能な基本レヤーまたはズーム・レヤーの場合

レヤーとその可視化表現の対を以後ビューと呼ぶことにする．2つの基本レヤーのビューが合成可能であるのは，両者のデータ型 (次元とその次元に対応づけられた属性の定義域) が合成可能な場合である．例えば，図2では，年別のエアコンの売上げのデータ集合と，「エアコン」というキーワードを含んだ日付を含むニュース記事の集合を，それぞれ (年，売上高)，(日付，関連度) という2次元の可視化表現に対応づける基本レヤーを，(年，売上高／関連度) という属性上に合成している．紙面の都合で合成可能性の形式的な定義にはふれないが，例のような可視化表現の合成が，一見比較不能な定義域をもつ可視化表現の場合や，次元の異なる場合なども含めて定義可能である．フィルター機能は，ビュー中の可視化表現との対話操作 (マウス・クリックなど) として実現される．

また，2つの2次元の可視化表現の1つの次元が同じであれば，その次元で両者を隣接させることにより相関関係を直観的に把握できる図2の例では，「年」と「日付」を年のレベルで共有することで，図3のように隣接表現することができる．

理論的には，1つの画面をmxn個のマトリクス状に分割するレヤーを考え，それぞれ隣接する可視化表現と次元を共有するようなmn個のデータ集合と可視化表現を考えることが可能である．このようなレイアウトは，データ・マイニング・ツールのように，様々な属性の相関関係の可視化よって新事実／法則を発見するような問題領域には有用であり，テキスト・マイニングにも同様の期待があるが，一般的には「exact cover by 3-sets」というNP-完全の問題 [2] を解くのと同じになるため効率的な計算は困難であると予想される．

3 応用分野

ここでは前節で示した情報可視化のモデルが実際にどのように応用分野に適用されるかについて考察する．

3.1 情報検索

情報検索で重要なのは，関連度フィードバックなどによって検索式を修正し，求める情報により関連した文書を高い精度で検索する場合 (特定検索) と，大まかな文書集合から特徴的なキーワードやクラスターをもとに，興味のある文書集合を絞り込む場合 (広域検索) の2種類の全く異なる情報要求を支援することである．このために文書集合を概要レベルで可視化するレヤーと，特定の文書内容から検索式を修正しやすいように詳細レベルで可視化するレヤーを設計する．具体的には，図4のように，文書集合が得られた場合に，L_1のように分野別／頻度順のキーワード分布や，月別の文書数の分布から広域検索を支援する．一方で，L_2のように特定の文書の内容に注目し，L_3でその要約や語の関係を可視化することで関連度フィードバックに用いる用語を決定する特定検索を支援する．この図では，L_3は分割レヤーによってテキスト情報を2つの可視化表現にしているが，ビュー合成を用いてテキスト文に直接付加情報を重ね合わせるよ

[1]ただしテキスト・マイニングのように，大量のテキスト情報をある程度意味的なレベルまで解析して，対話的な意思決定を支援するという大規模計算を行う応用もある．

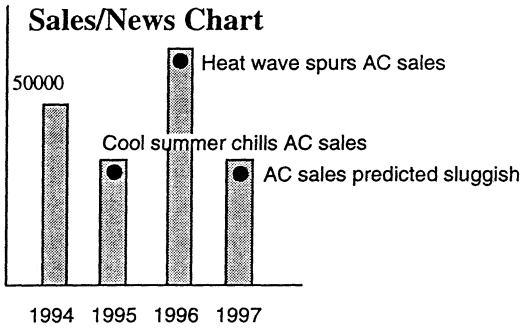

図 2: エアコンの売上げとニュース記事の合成表示

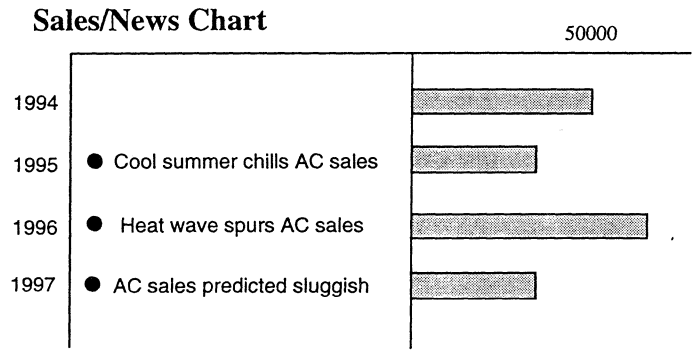

図 3: エアコンの売上げとニュース記事の隣接表示

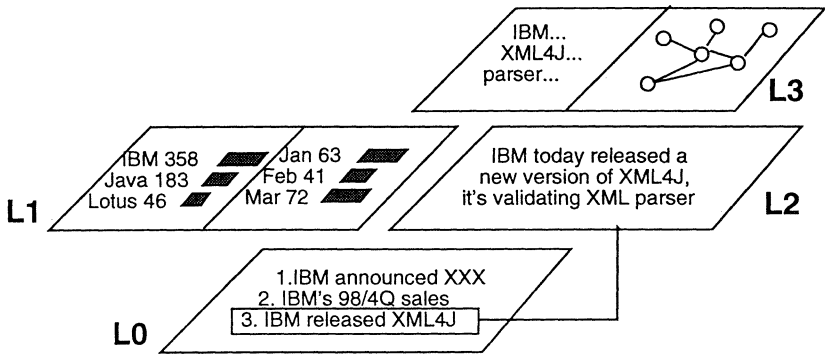

図 4: レヤーを用いた情報検索のための可視化

うなこともできる．これは，いわゆる注釈(annotation)の技術[4]で，テキスト情報だけを，より実世界のメタファに近い形で提示できる．

また，ここで重要なことは，これらのレヤー上を視点の位置(レヤー上の位置)を変えながら移動することで，両方の検索方法が混在したような情報検索が可能になることである．従来の可視化では，ビューの重ね合わせやレヤーのキュー上での視点の移動という概念が陽に意識されていなかったため，可視化表現がユーザとの対話に応じて不連続に変化するという使いにくさがあった．本モデルでは，ユーザの視点の変化とそれに伴って増減する可視化情報の変化がキュー上で確認できるとともに，これもナビゲーション機能の一部として制御可能になっている．レヤーのキュー中に存在するズーム・レヤーは，可視化のレベルが変化したことに対応するので，これを適切に管理することで，ユーザに適用可能なビューのメニューを提示したり，これらのレベル間を順にナビゲートすることが容易になる．

3.2 情報分析と意思決定

従来の情報分析では，データベースのように定型化されたデータを各種の統計手法を利用して，できるだけ多くのデータに高い確度で成立する性質を計算することが重要であった．データ処理能力の増加とともにスケーラビリティが要求され，計算の複雑さによって適用可能なデータ集合のサイズと分析手法とが制限される．従って構文解析のように文のサイズの3乗に比例するような計算は，たとえオフラインで実行しても利用困難であると考えられていた．しかし，テキストに含まれるモダリティ情報[9]や特定の自立語は，その文書の柔軟な分類に決定的な役割を果たすことがあり，いわゆるカタログデータ的な情報と相補的に利用するメリットは無視できない．情報分析で多用されるグラフ，クラスタ，決定木といった構造に，テキストから得られた情報を合成することで，ユーザの経験的な判断を大幅に効率化することを目標として，浅い構文解析やテンプレートによる情報抽出と，データ・マイニング技術を統合することが重要である．図2で示したようなビュー合成とレヤー構造は，このような意思決定を補助するとともに，データ集合のサイズに応じて計算時間の違うマクロ的およびミクロ的な情報分析を切り替える新しい技術として有望である．

4 あとがき

本論文では，自然言語処理と可視化の様々な技術が，対象分野とタスクの異なる多くの問題を解決するための共通のモデルとして統合できることを主張してきた．これらの技術は，Information Outlining[5]および site outlining[8]といったプロトタイプによって検証されつつあり，ここで提案しているビュー合成の実現とともに，テキスト・マイニングや知識管理といったより高度な応用分野での実用化を目指している．

参考文献

[1] J. W. Cooper and R. J. Byrd. "Lexical Navigation: Visually Prompted Query Expansion and Refinement". In *Proc. of ACM Conf. on Digital Libraries'97*, pages 237–246, 1997.

[2] M. Garey and D. Johnson. *"Computers and Intractability"*. W.H.Freeman and Co., San Francisco, 1979.

[3] X Lin. "A Self-Organizing Semantic Map for Information Retrieval". In *Proc. of SIGIR'91*, pages 262–269, 1991.

[4] C. C. Marshall. "Annotation: from paper books to digital library". In *Proc. of ACM Conf. on Digital Libraries '97*, pages 131–141, 1997.

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

[6] 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.

[7] B. Shneiderman. *"Designing the User Interface: Strategies for Effective Human-Computer Interaction" (3rd Edition)*. Addison-Wesley, 1998.

[8] K. Takeda and H. Nomiyama. "Site Outlining". In *Proc. of ACM Digital Libraries'98*, pages 309–310, Pittsburgh, PA., Jun. 1998.

[9] 乾, 内元, 井佐原: "モダリティ分析に基づく自由回答アンケートの分類". 言語処理学会 第*4*回年次大会, pp.540–543, 1998年.

[10] 那須川, 諸橋, 長野: "テキスト マイニング: 膨大な文書データからの知識獲得". 情報処理学会第*57*回全国大会, 5K–4, 1998年.