1D3-2

情報編纂（Information Compilation）の基盤技術

Toward Information Compilation

加藤 恒昭[*1] 松下 光範[*2]
Tsuneaki Kato Mitsunori Matsushita

[*1]東京大学 The University of Tokyo
[*2]日本電信電話株式会社 Nippon Telegraph and Telephone Corp.

Information compilation is a novel technology for tackling information flood. It handles linguistic and non-linguistic information and generates its "summary" that allows users to understand and utilize that information. That summary could be a multi-media presentation that summarizes underlying information or an interactive interface mechanism for accessing it. In this paper, we characterize a new research field on information compilation and describe our first attempts toward it.

1. はじめに

情報洪水という言葉が既に古びたものと感じられるほど，情報過多は日常的な状況となり，社会はそれへの対応を迫られている．それらの情報は，文章，図表，グラフィクス等，様々なメディアによって表現され，その構成や形式も多様である．このような状況での情報の利用には本来は不要であるはずの様々なスキルが要求され，それを持たない人達が不利益を被るという情報格差の存在もその指摘があって久しい．言うまでもなく，豊富な情報の存在はそれを利用する適切な手段が提供されれば歓迎すべき状況である．多くの人が利用でき理解できる情報であればそこに格差は生じない．本稿では，雑多な情報を知的に編纂し，それらの理解を容易にすると共にそれらへの簡明なアクセス手段を提供するための基盤技術を情報編纂–Information Compilation–と名付け，その研究開発を通じて，情報洪水を情報格差のない情報豊富という望ましい状況へと向かわせることを提案する．本稿の構成は以下の通り．まず，情報編纂がどのような技術であり，既存の研究分野とどう関係するかを説明する．その後，筆者らが行ってきた情報編纂へと向けた研究について報告し，続いて，関連する研究動向を報告しつつ，情報編纂基盤技術の今後を展望する．

2. 位置づけ

情報編纂基盤技術とは，情報の理解とそれへのアクセスを支援するシステムを構築する基盤技術である．そこでは，利用者の関心に基づいて多量のテキスト等の言語情報がその意味内容の理解を通じて洗練・要約されると同時に，同じく利用者の関心に基づいて数値データや図表等の非言語情報が解釈され，そこから必要部分を抜き出す，様式を変換する，その内容を文章で表現する等のメディア理解が行われ，それらによって得られたものを協調的に組織化することで情報内容の概観を可能とする総合的な要約が生成される，あるいは，それらを通じてその背後にある情報へのインタラクティブなアクセスを可能とする仕組みが作り出される．

この情報編纂技術は，テキストを中心とする言語情報と数値情報や視覚情報等の非言語情報とを利用者の関心に応じてマルチメディアプレゼンテーションとしてまとめあげるマルチモーダル要約技術と捉えることもできる．従来のマルチメディアプレゼンテーション生成の技術 [2] が既に組織化された整った情報を対象としそれに関する知識の存在を前提とした技術であったのに対し，マルチモーダル要約としての情報編纂は現実に存在する組織化されていない情報から，その内容を概観できるプレゼンテーションあるいはそれらの情報へのインタラクティブなアクセスを可能とする仕組みを作り出す点が特徴的である．これらの関係は，言語生成技術が整った意味表現から言語情報を生成するのに対し，テキスト要約が現実に存在する組織化されていない文書からその内容をまとめあげる技術であるという言語生成とテキスト要約との関係に等しい．ここで，テキスト要約においては，原文の代わりに用いられそれだけで内容を理解する報知的（informative）要約と，原文が読むに値するかの判断等，原文を参照する前の段階で用いる指示的（indicative）要約とが分けて論じられるが，マルチモーダル要約においても同様の分類を考えることができ，それ自体で情報の概観が可能となるマルチメディアプレゼンテーションの生成は報知的要約に，必要部分へのアクセスを容易にするための言語的視覚的な情報の提示とインタラクティブな情報アクセスの支援は指示的要約に対応するといえる．

関連する研究は様々な形で行われてきているが，入力や出力として言語情報を対象とするか非言語情報を対象とするか，報知的な要約か指示的な要約か，つまり，それだけで材料となった情報の内容を概観する新しい情報を作成するか，背後にある情報にアクセスするインタフェースを構築するかで，異なる技術とされ，異なる研究コミュニティに属している．

まず，言語情報を対象にその言語的な要約を生成するのはテキスト要約技術 [15] として研究されている．報知的な要約に加えて指示的な要約も扱われているが，指示的な要約を使った情報アクセスの仕組み全体は要約研究の外側にあると考えられるのが一般的である．例えば，そういう仕組み全体に関連する評価は extrinsic な評価ということで "要約そのもの" に関する intrinsic な評価と区別される．テキスト要約のインタラクティブなインタフェース構築の例 [7] もあるが，情報アクセスの仕組み全体の研究はむしろ文書可視化ということで CHI や情報可視化，情報検索の分野で扱われ，文書を時間軸と内容の軸からなる空間に配置し，魚眼レンズを模して焦点と文脈を表示する Perspective Wall[9] をはじめとして，その文書が関連する時間と場所に基づいた配置 [21], 文書のトピックやキーワードの相互関係に基づく配置 [20] 等とそれらを用いた情報アクセスが研究されている．

一方で，非言語情報，ここでは後述する関心から主に数値データ，グラフ，表を考えるが，については，数値データと関連した自然言語生成，つまり非言語情報の言語情報による報知的要約が，レポート生成として長く研究が続けられている

[5]．非言語情報の非言語情報による要約については，シミュレーションデータ等の多量の数値データを可視化するいわゆる科学的可視化 [13] をそのような要約ととらえることができるし，数値データへのインタラクティブなアクセスの支援では，情報可視化として分類される Table Lens[17] や Dynamic Queries[19] 等，古くから多くの研究がある．また，マルチメディアプレゼンテーション生成の分野では，数値データをその特徴，利用の意図，表現の意図等を考慮して，適切なグラフとして提示する研究が続けられている [1][2]．

言語情報と非言語情報という区別から離れ，テキストやグラフの素材となるような事実やデータの集まりを抽出し出力とする技術も研究されている．言語情報であるテキストを解析して事前に指定された情報を抜き出す技術が情報抽出として研究されており [18]，その意味論が事前に与えられていない表等を理解してその内容である情報やデータを抽出する技術は Web ラッパの構築と関連して研究されている [23]．

情報編纂の基盤技術の研究では，これらすべてを情報の理解と利用を促進する知的な編纂という統一的な視点のもとに整理し，個別のアプリケーションにとどまらず，特定の研究分野にこだわらない新しい技術を研究していく．言語情報と非言語情報を広く扱い，その両方を視野に入れることで，それらの相互作用や協調の可能性を追求することが特徴であり，従来のマルチメディアプレゼンテーション生成の研究とは違い，現実に存在する雑多な情報を編纂し，要約するという点が独特である．

3. これまでの研究成果

前節で見てきたように情報編纂を構成する技術は様々な観点で分類することができ，それらは有機的に関連しているが，以下では図 1 に示すような構成を仮定する．ここでは処理の対象となる入力の分類に基づいて構成要素を定義している．言語理解技術，メディア理解技術はそれぞれ言語情報，非言語情報を対象とし，利用者の関心を受けて，それを理解し，取捨選択し，要約し，様々な様式（メディア）からなる情報単位を生成する．プレゼンテーション・インタフェース生成技術はそれらの情報単位を協調的に組織化することで情報内容の概観を可能とする総合的なレポート，あるいは，それを通じてその背後にある情報へのインタラクティブなアクセスを可能とする仕組みを生成する．

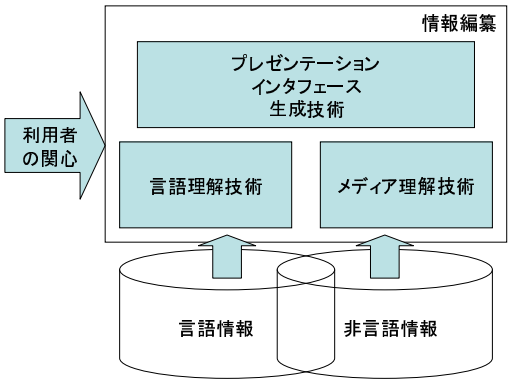

図 1: 情報編纂技術の構成

筆者らは，情報編纂に関わる，あるいは情報編纂という枠組みに思い至らせたふたつの研究を既に進めている．

第一は，グラフを用いて対話的探索的なデータ分析を支援するシステム InTREND[10][12] である．このシステムは背後にあるデータとしては数値データを扱い，メディア理解技術とプレゼンテーション・インタフェース生成技術に関連する．

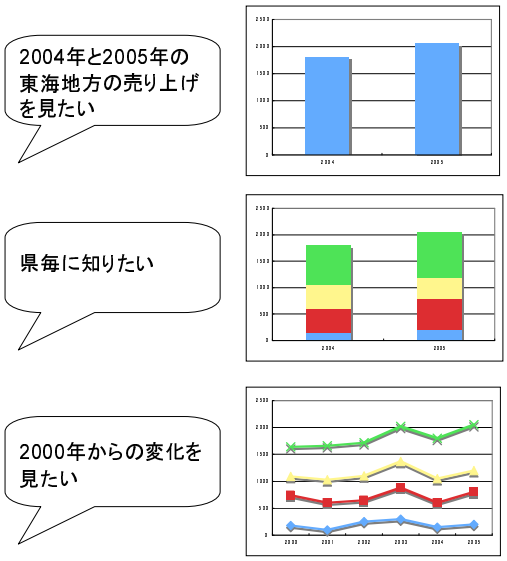

図 2: InTREND による対話

InTREND は，図 2 に示すように，利用者の一連の発話に対して，数値データの適切な部分を，適切な形式のグラフに描くことで応答する．グラフの形式は棒グラフ，折れ線グラフ，積み上げ棒グラフ等から選択される．言語情報によって表現された関心に従って数値情報を編纂して提示する点が特徴で，直接操作を中心とした従来のアクセス支援と区別される．自然言語発話という言語情報を使用することでデータの探索的分析というタスクの言語レベルの思考がそのまま計算機に伝えられることに加えて，対話ということで，利用者は代名詞や省略を意識することになく自然に利用し，それを通じて自分の意図を伝えることができる．例えば，図の対話において第二の発話の応答となるグラフは，描かれるデータは「東海地方の各県の 2004 年と 2005 年の売り上げを見たい」という関心に応答するグラフと同じであるが，利用されるグラフの形式は異なる．図の対話では，関東地方の売り上げ全体に関する関心がそれ以前に示されているのでそれに応じたグラフが選択されるためである [3]．このように，InTREND システムでは，言語を用いた人間どうしの対話と同様に意図や関心の伝達を自然に行うことができ，それを通して膨大な数値データを対話的探索的に分析できる．

第二は，現在研究を進めている動向情報の可視化システム STEND である．このシステムは，利用者の関心となっているトピック，例えば，最近のガソリンや原油の価格動向，携帯電話の普及状況等を入力とし，新聞記事集合からそれに関連する情報を収集し，そこから得られる動向をグラフとして提示し，利用者がそれを通じて新聞記事情報にアクセスすることを許している*1．新聞記事を情報源とするということで，言語理解技術とプレゼンテーション・インタフェース生成技術に関連する．

テキストからグラフ描画に必要な情報を抽出する部分はいわゆる情報抽出技術に近いが，ふたつの点が特徴的である [11]．まず，同じトピックに関連する一連のテキストから一連の情報を抽出するために，具体的直接的に表現された情報だけでな

*1 このシステムは後述する MuST ワークショップのデータセットを利用している．また現時点ですべての実装が完了しているわけではない．

く,「昨年 10 月より約 40%の下落になっている」「前年同月に比べて 5 ドル上昇した」等の比較表現も利用して情報を抽出する．次に，具体的な数値データだけでなく,「安定傾向にあった」「10 月をピークに下落している」等の定性的表現，蓋然的表現も情報として抽出する．特に後者の情報はそのトピックについて単なる数値の羅列にとどまらないいわゆる動向を抽出するものである．従って描かれるグラフも抽出されたデータの点を結んだだけのものでなく，抽出された動向が「ピーク」「上昇」「安定」等，グラフ概形の基本パタンと対応づけられ，その概形を表現するアイコニックな図形（グラフプリミティブと呼んでいる）がグラフに貼付けられる．図 3 にその様子を示す．点が比較表現等の利用を含めた情報抽出で得られたデータ，様々な矢印記号が定性表現から得られたグラフプリミティブである．矩形は値や日時が蓋然的であるような情報に対応している．グラフ上のこれらの点やグラフプリミティブ等を指定することで利用者はその情報が含まれている新聞記事にアクセスすることができる．

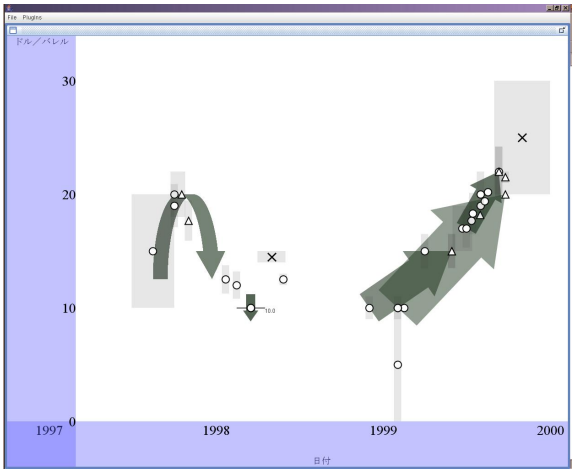

図 3: STEND の出力例

InTREND は数値情報に言語的なインタフェースを与え，それをグラフという視覚情報で提示する．STEND は言語情報を要約し，それに視覚的なインタフェースを与える．更に STEND では，数値情報と言語的な情報を融合し，ひとつの視覚的枠組みで提示することを試みている．このようにこれらのシステムは異なるメディアの相互作用や協調の有効性を示しており，それを活かした情報アクセス，情報理解の大きな可能性を示唆するものである．

4. 研究のひろがりと今後の展開

情報編纂は，言語情報と非言語情報を広く扱うことをその特徴のひとつとするが，様々な非言語情報の中で「動向」としてまとめられるような統計量等の数値データ，出来事に関するデータは，言語情報との協調の幅が広く，情報編纂基盤技術の初期検討としてそれらを対象とするのが適切であると考えている．前節で述べた InTREND, STEND の成功もその一例となっている．そのような関心を動機のひとつとして，筆者らは「MuST:動向情報の要約と可視化に関するワークショップ」を運営すると共に、そこに参加している [4]．MuST ワークショップでは，ガソリン価格，通信機器，パソコン，地震，台風など 20 のトピックに関連した新聞記事集合を動向情報の抽出という観点で意味的に註釈した共通のデータセットを用いて，動向情報の要約と可視化という緩い意味で共通したテーマについて参加者が競争的かつ協調的に研究が進められている．本年 3 月には第一回の成果進捗報告会が開催された [14]．

このワークショップの枠組みの中から，情報編纂の基盤技術に繋がる多くの技術が生まれている．山本らは STEND と同様にテキスト情報からグラフ描画に必要な数値データを抽出し，それを使って描かれたグラフを通じた情報アクセスインタフェースを提供しているが，そのグラフでは，そこに描かれた統計量が特徴的な変化を見せた時点にその変化の原因となった要因の候補が注釈づけられている [25]．このような注釈付けは STEND とは違ったグラフと言語情報の協調であり，グラフを単なるデータの提示以上のものとしている．小林らの一連の研究では，日経平均株価の情報を用いて，グラフの挙動を説明するテキストの生成やグラフの詳細度と協調したテキストの要約等，数値データと言語情報とのメディア変換やメディア協調が研究されている [16][22]．山田らは，地震の発生と規模に関する情報を扱い，時間的空間的な広がりを持つ出来事情報に対して，空間的時間的動向の抽出とその可視化を通じた情報アクセスインタフェースを提供している [24]．また，彼らは地震情報のように詳細な数値データがある状況に新聞記事の情報を組み合わせることの意義として，新聞記事が出来事の重要さ等に関するフィルタになること，地震による交通機関の遅延等，関連する出来事へのリンクを提供することを指摘している．

この他にも，テキストからの情報抽出を中心とした言語情報の要約等，幾つかの研究が進められている．今後も，これらの研究を積極的に進めるとともに基盤技術として整理していくことが必要である．また，様々なシステムや提案の一般化と共通項の括り出しが，MuST ワークショップを情報編纂基盤技術の研究へとつなげていくための課題である．

情報編纂基盤技術の研究は，まず，数値データと言語情報を動向としてまとめ，その動向を言語的視覚的に表現し，それを通じてその背後にある詳細な情報への柔軟なアクセスを可能とするという，STEND を例とするような枠組みを通じて進めていき，その基盤技術を明らかにしていく．この枠組みでは，MuST ワークショップがそうであるように，動向情報としてその背後に時間的空間的な数値データや出来事に関するデータが存在するものを中心に扱っていく．一方で，動向という概念はそれ以上の広がりを持っている．現在積極的に研究が進められている評価，評判，態度，意見等に関する情報抽出[*2]とそれらの可視化 [8][6] は，広い意味での動向，いわゆるトレンドを要約理解するものとして情報編纂技術の一部を構成し，次なる展開の最右翼である．更に，画像理解や画像要約，音声や動画等，非言語情報の検索メカニズムの構築も情報編纂という枠組みの中で整理でき，統一的な視点で研究が進められると確信している．

5. おわりに

言語情報と非言語情報を広く扱い，その理解と利用を支援するための技術として情報編纂を提案した．その基盤技術について，数値データと言語情報を動向としてまとめ，その動向を言語的視覚的に表現し，それを通じてその背後にある詳細な情報への柔軟なアクセスを可能するという枠組みの構築を最初の目標として研究を進めていく．情報編纂はその対象，その技術に様々な広がりを持ち，それらに統一的な視点を与えるものである．

*2 例えば言語処理学会第 12 回年次大会ワークショップ「感情・評価・態度と言語」で発表された一連の研究等.

謝辞

MuST は本稿著者らに加えて国立情報学研究所神門典子氏によって運営されています．本稿の内容について MuST 参加者から様々な示唆を頂いています．これらの皆様に深く感謝いたします．また，要約と可視化のサーベイについては，NTCIR-5 Workshop Meeting での Dr. Chin-Yew Lin の講演を参考にさせていただいています．あわせて感謝いたします．本研究の一部は文部科学省科学研究費 (課題番号: 17700168) の助成を受けています．

参考文献

[1] M. Fasciano and G. Lapalme. “Postgraphe: A system for the generation of statistical graphics and text”, *Proc. 8th International Workshop on Natural Language Generation*, pp. 51 – 60, 1996.

[2] 加藤 恒昭.「マルチメディアプレゼンテーションの自動生成に向けて－自然言語生成からマルチメディア生成へ－」, 情報処理, Vol. 38, No. 12, pp. 1049 – 1056, 1997.

[3] 加藤 恒昭，松下 光範.「グラフを用いた探索的データ分析のためのマルチモーダル対話処理」，電子情報通信学会論文誌 D-II, Vol. J87-D-II, No. 5, pp. 1142 – 1152, 2004.

[4] T. Kato, M. Matsushita, and N. Kando. “MuST: A Workshop on Multimodal Summarization for Trend Information”, *Proc. NTCIR-5 Workshop Meeting*, pp. 556 – 563, 2005.

[5] R. I. Kittredge and A. Polguère. “The Generation of Reports from Databases”, R. Dale, H. Moisl, and H. Somers eds. Handbook of Natural Language Processing, pp. 261 – 304, Marcel Dekker, Inc. 2000.

[6] B. Lee, M. Czerwinski, G. Robertson, and B. B. Bederson. “Understanding Research Trends in Conferences using PaperLens”, *Extended Abstsracts of CHI 2005*, pp. 1969 – 1972, www.cs.umd.edu/hcil/paperlens/, 2005

[7] A. Leuski, C. Lin, and E. Hovy. “iNeATS: Interactive multi-document summarization”, *Proc. the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pp. 125 – 128, 2003.

[8] B. Liu, M. Hu, and J. Cheng. “Opinion Observer: Analyzing and Comparing Opinions on the Web”, *Procs. The 14th International Wirld Wide Web Conference (WWW2005)*, 2005.

[9] J. Mackinlay, R. Rao, and S. K. Card. “The Perspective Wall: Detail and Context Smoothly Integrated”, *Proc. of the ACM Cnference on Human Factors in Computing Systems*, pp. 173 – 179, 1991.

[10] 松下 光範，中小路 久美代，山本 泰裕，加藤 恒昭.「探索的データ分析における “自然なやりとり” の実現に向けて — インタラクティブ可視化システム InTREND –」情報処理学会　インタラクション 2003，pp. 99 – 106，2003.

[11] 松下 光範，加藤 恒昭.「動向情報テキストに基づく統計グラフ描画方式の検討」，電子情報通信学会 NLC 研究会「テキスト情報の要約と提示に関わる自然言語処理シンポジウム」予稿集，pp. 25 – 30, 2006.

[12] 松下 光範，加藤 恒昭.「コンテキスト保持による探索的データ分析支援の枠組」，知能と情報, Vol. 18, No. 2, (印刷中), 2006.

[13] B. H. McCormick., T. A. DeFanti, and M. D. Brown. “Visualization in Scientific Computing”, Computer Graphics, Vol. 21, No. 6, pp.1 – 14, 1987.

[14] http://must.c.u-tokyo.ac.jp

[15] 奥村 学.「テキスト自動要約」, 情報処理, Vol. 45, No. 6, pp. 574 – 579, 2004.

[16] 奥村 奈穂子，小林 一郎.「グラフの挙動を表すテキスト生成」，言語処理学会 第 12 回年次大会 ワークショップ「言語処理と情報可視化の接点」予稿集，pp. 17 – 18, 2006.

[17] R. Rao and S. K. Card. “The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information”, *Proc. the ACM SIGCHI COnference on Human Factors in Computing Systems*, pp. 318 – 322, 1994.

[18] 関根 聡.「情報抽出－情報を整理して提示する－」， 情報処理, Vol. 45, No. 6, pp. 563 – 568, 2004.

[19] B. Shneiderman. “Dynamic Queries for Visual Information Seeking”, IEEE Sofrware, Vol. 11, No. 6, pp. 70 – 77, 1994.

[20] A. Takano, et al. “Associative Information Access Using DualNAVI”, *Proc. Kyoto International Conference on Digital Libraries (ICDL’00)*, pp.285 – 289, 2000.

[21] 武田 浩一，野美山 浩.「テキスト情報の可視化を利用した情報検索」，情報処理, Vol. 41, No. 4, pp. 343 – 350, 2000.

[22] 渡邉 千明，小林 一郎.「グラフと協調するテキスト要約」，言語処理学会 第 12 回年次大会 ワークショップ「言語処理と情報可視化の接点」予稿集，pp. 19 – 21, 2006.

[23] 山田 泰寛，池田 大輔, 他.「WWW からの情報抽出－ Web ラッパーの自動構築－」，人工知能学会学会誌，Vol. 19, No. 3, pp. 296 – 301, 2004.

[24] 山田 隆志，中野 純，高間 康史.「タグ付きコーパスを用いた地震記事からの地理的動向情報可視化」，言語処理学会 第 12 回年次大会 ワークショップ「言語処理と情報可視化の接点」予稿集，pp. 9 – 12, 2006.

[25] 山本 健一，殿井 加代子，谷岡 広樹.「タグ付きコーパスを用いた動向情報とその要因の可視化」，言語処理学会 第 12 回年次大会 ワークショップ「言語処理と情報可視化の接点」予稿集，pp. 13 – 16, 2006.