

情報編纂研究会が目指すもの

The Goal of Special Interest Group on Information Compilation

加藤 恒昭*1

Tsuneaki Kato

*1東京大学

The University of Tokyo

Information compilation is a technology for compiling variety of information and making it easy to understand. It supports interactive information exploration, in which users try to find information they need and forge it into knowledge. The objectives of Special Interest Group on Information Compilation (SIG-IC) include defining the framework of information compilation, constructing its technological infrastructure, and proposing its evaluation measure. This paper describes what is information compilation and what roles SIG-IC plays for information compilation in some detail. It also introduces Interactive Visual Exploration Task (VisEx) of an evaluation workshop series, NTCIR, which shares with SIG-IC its interests on the evaluation measure.

1. はじめに

情報編纂は、様々な情報を編集・編纂してわかりやすくするための技術である。ただし、編集・編纂一般ではなく、利用者が必要な情報を見つけ出し、それを知識に変えていくための対話的・探索的な情報アクセスを支援することに焦点をあてている。わかりやすく示すための手段として視覚情報の利用に重きをおいている点、対象とする情報として、テキスト等の言語情報だけでなく、数値情報等、様々なメディアを視野に入れている点も特徴である。

情報編纂研究会は、この情報編纂技術に対して、研究分野としての特徴づけ、インフラの構築、評価の枠組みの提案等をすすめることを目的として活動している。その一環として現在は、関連分野のチュートリアルや情報編纂に関わる研究報告を中心としたシンポジウムを開催している。そして、情報アクセス技術の研究促進を目的とした評価ワークショップ・シリーズである NTCIR に、情報可視化を用いた対話的探索の評価方法を検討するためのパイロット・タスク VisEx を提案し、そこでの活動を通じて、評価の枠組みの検討やインフラの構築を進めている。

本稿では、まず、対話的・探索的な情報アクセスの過程がどのようなものであるかを描くことで、それを支援する情報編纂の技術を位置づける。そして、これまでになされた提案を例示することで、システムの具体的なイメージを示す。次に、情報編纂研究会が情報編纂技術に対してどのような観点で寄与しようとしているかを述べ、その活動を紹介する。最後に、VisEx についてその着眼と概要を紹介する。

2. 情報編纂とは

2.1 対話的・探索的な情報アクセス

対話的・探索的な情報アクセスは、必要な情報を見つけ出し、それを知識に変えていくという活動である。自分が関心を持つ経済や社会の状況の動向を要約したい、ある期間の特徴的な変化や出来事を調べて、その影響や意味づけを明らかにしたい等と利用者が考えることがその動機や目的となる。

その中で最初に行なわれることは、利用者の関心を表わす

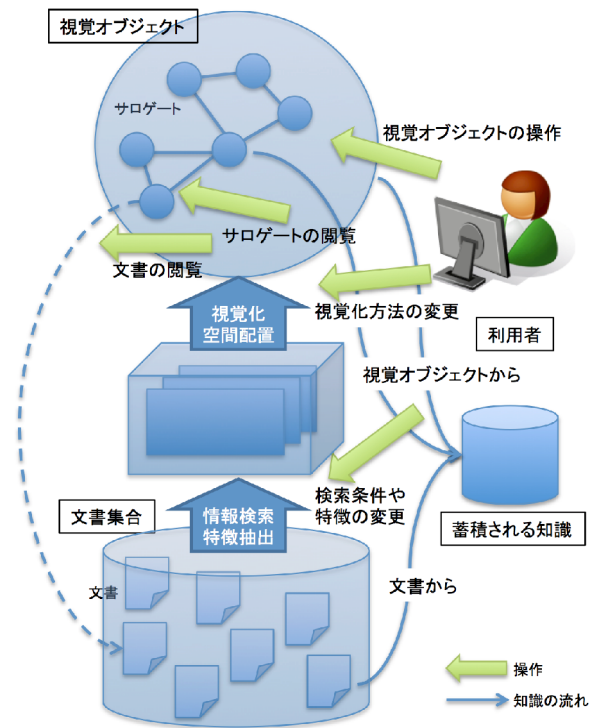


図 1: 対話的・探索的な情報アクセスと視覚オブジェクト

キーワード等による、対象となる情報の取捨選択である。ここで、対象となる情報の単位を文書と呼ぶ。web ページや新聞記事もしくはそのパッセージに加え、数値情報の表、画像等が文書となりうる。文書の取捨選択はいわゆる情報検索の手法によって行われる。そして、選択された文書が編集・編纂され利用者に提示される。ここでは、選択された文書が解析され、特徴の抽出が行われ、その概要を表現するサロゲートが生成される。そして特徴に基づいてサロゲートが配置され、視覚オブジェクトの構築、視覚化が行われる。この様子を図 1 に示す。

サロゲートは、文書の概要や特徴を表現すると同時に、その文書にアクセスする際の入り口となるような視覚的な手掛かりあるいはアイコンである。視覚オブジェクトは広い意味での空

間にサロゲートが配置された視覚的実体で、概観や操作の対象となる。視覚オブジェクト内でのサロゲートの位置も文書の特徴や文書の間の関係を表現する。逆に言えば、サロゲートを作成しそれを配置した視覚オブジェクトを構築するために、必要な文書の解析と特徴抽出が行われる。一般の web サーチエンジンが出力する検索結果一覧は、テキストを中心としたページであるが、これが視覚オブジェクトの簡単な例である。文書の適合度が特徴量であり、タイトルとスニペットがサロゲートとなる。そして、文書の適合度を基準としてサロゲートが一次的に配置されている。

より有益な視覚オブジェクトとして以下のようなものが考えられる。

- 地震の発生のような出来事に関する情報を探している場合に、それが発生した実際の地理的時空間にその出来事に関するサロゲートを配置することで、視覚オブジェクトとする。文書中から出来事の発生した地点や時点を抽出することが特徴抽出となる。
- 政治状況における内閣支持率のように何かの数値情報が関連している場合に、その数値情報のグラフにサロゲートを配置して視覚オブジェクトとする。グラフはそれ自体が情報として存在する場合もあれば、文書の内容から作成されることもある。配置される文書が関連するグラフの部分を明らかにすることが特徴抽出となる。
- 主題の類似性等を距離と関連づけた超空間を視覚化可能な空間に写像して、そこにサロゲートを配置したものを視覚オブジェクトとする。個々の文書だけでなく、クラスタリングの結果が配置されることもありうる。
- リンクで繋がれた web ページのように文書集合がネットワーク構造を持つ場合に、そのトポロジを反映した空間にサロゲートを配置して、視覚オブジェクトとする。
- 文書が数値情報の場合、それを一定の粒度でグラフ化して表現すると共に、その値や変化を文章によって要約したものをサロゲートとして配置することで、視覚オブジェクトとする。

これらの視覚オブジェクトとそこに配置されたサロゲートを通じて、利用者は情報を概観し、その傾向を把握していく。それらは文書の内容を容易に把握し理解することを助け、それがそもそもどんな情報であり、必要であるのか、全体や今着目されている部分の中でどんな位置づけにあるのかを少ない負荷で判断することを可能とする。そして、どのような情報がどこにあるのかという全体の傾向を伝える。ある時期に大きな変化が起こっておりそれに関する情報が多いことがわかるかもしれない。特定の場所に出来事が集中していることが明らかになるかもしれない。そのような発見が特定の詳細への興味の変化を引き起こすことになる。逆にその時点で得られているものは期待していた情報ではないことが明らかになるかもしれない。利用者が求めているものは、利用者本人にとっても常に正しく意識されているとは限らないし、探索の過程において変化し発展するかもしれないものである。そのような意識化や変化を引き起こすのもこの視覚オブジェクトである。

この視覚オブジェクトに対して、概観や傾向の把握のための操作や、関心の変化や具体化詳細化に伴う操作が行われる。それらは図 1 のモデルに即せば以下のように分類・実現される。

- 視覚オブジェクトの操作。地図やグラフのズームングやパニング、空間に対する視点の変更等。

- サロゲートの閲覧（ブラウジングやスキニング）
- 個々の文書の閲覧
- 視覚化に用いる特徴の変更や軸を替えての視覚化。空間的な配置を時間的な配置に変更する。適合度順に並んでいた文書を発行日順に並び替える等
- いわゆる情報検索的な情報要求の変更や絞り込み。明示的な検索条件の変更に加え、適合性フィードバックや関連検索によるもの等

概観や発見のための操作、それに続く操作を許すという点で、視覚オブジェクトは情報を表現するプレゼンテーションであると同時に、操作される対話的オブジェクトであり、情報アクセスのためのインタフェースでもある。

視覚オブジェクトの観察から得られる傾向や、アクセスされた文書の内容が、利用者が獲得していく知識の素材となる。利用者は、それらを分析・比較し、それらを抜き出したり集約したりして、知識として構成していく。この分析や比較は十分な時間がかけられた文書の熟読による場合も、足早なスキニングによる場合もある。得られた知識によって必要と判断される情報が変化することもありえるし、今後の検索に利用できる新しいキーワードが得られることもある。それらも視覚オブジェクトへの新たな操作のきっかけとなる。

このような過程を繰り返す中で徐々に知識が組織化されていく。必要な情報を見つけ出し、それを知識に変えていくための対話的・探索的な情報アクセスは、このような変化や発展を伴う累積的な過程である。Bates はこれを木イチゴ摘み (Berry-Picking) に例えている [2]。

2.2 目的と位置づけ

以上に示したような対話的・探索的な情報アクセスの過程全体を一貫して支援しようとするのが、情報編纂である。その中でも様々な情報を編集・編纂してわかりやすくするために重要となるのが、視覚オブジェクトの生成である。そのための特徴抽出やサロゲート生成には情報抽出やテキストマイニング等の自然言語処理技術が中心となり、それに情報可視化技術が加わって視覚オブジェクトの構築、視覚化が行われる。これらは相互に関連し、言語情報と視覚情報の両方に関わるので、全体として、マルチモーダル要約、マルチメディアプレゼンテーション生成として位置づけることもできる。そして、この視覚オブジェクトが対話的オブジェクトであり、様々な情報アクセスの手段と結びついたインタフェースでもあることから、インタラクション・デザインや対話的な情報検索と関連する。

これらの技術はそれぞれ活発に議論されているが、情報をわかりやすくする、対話的・探索的な情報アクセスを支援する、更に進んで、そのための環境を提供するという視点でそれらを包括的に眺めようとする点に情報編纂技術の意味がある。また、言語情報や視覚情報をはじめとする複数メディアを活用すると同時に、処理の対象としてもそれら全体を意識するという点も特徴である。

例えば、図 2 に示すのは、エネルギー事情の調査のために、原油価格のグラフに文書を関連づけた視覚オブジェクトである。2つの画面のグラフにおいて丸点や矩形で示されているのがサロゲートとなる。下の画面ではグラフ上のサロゲートに対応するスニペットが右側に列挙されている。一方を選択すると他方の対応する部分がハイライトする等の対応づけがなされる。

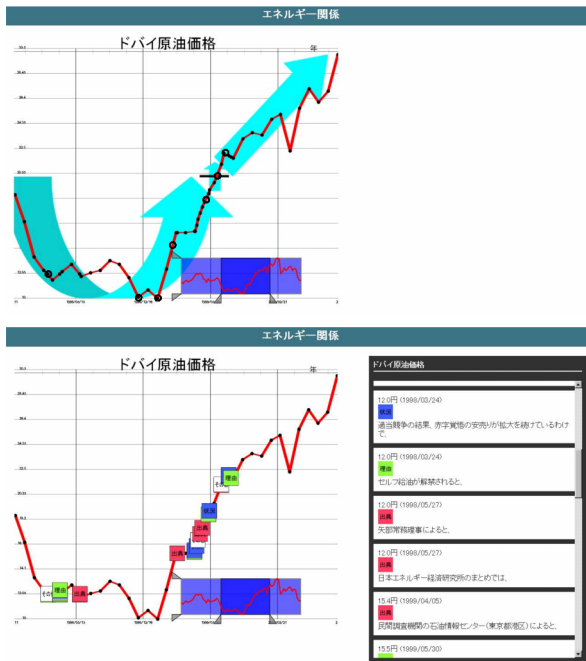


図 2: 文書とグラフを関連づけた視覚オブジェクト

このような視覚オブジェクトの生成のために、まず、文書に対して、ドバイ原油価格等の統計量を中心に特徴抽出がなされる。以下のような処理が必要である。

- ある記述が着目している統計量についてのものであることの認識（同じ統計量が様々な表現で言及されることへの対処）[7]
- その記述がどの時点の値がどれほどであると述べているかの認識 [8]
- 「高騰が続いている」等、より定性的な記述の抽出 [5]
- 統計量を取り巻く状況、例えば、変化の原因や影響、今後の予測等の分類と抽出

これらの特徴が視覚化される。ここでは、定性的な記述を矢印等のアイコンによって（図 2 上画面）、状況記述の分類を色分けを用いて（同下画面）、表現している。

視覚オブジェクトの操作として、グラフ表示を操作するためのパネルがグラフの右下に配置されている。これを使って、表示する期間の変更（パニング）や、拡大縮小（ズーム）が行える。出来事のように時間情報に加えて場所情報を持つものについては、時間軸を持つグラフと地理的な情報を表わす地図とのそれぞれに対する操作に加えて、それらを関係づけ、行き来する操作を許さなければならない [9]。そしてこれらの視覚オブジェクトを介してサロゲートや文書の閲覧が行われる必要がある。これら全般を扱うインタラクションの設計が必要となる。

このような様々な構成要素を持つ技術を、様々な情報を編集・編纂して、わかりやすくする、必要な情報を見つけ出し、それを知識に変えていく対話的・探索的な情報アクセスの過程を支援する、という観点でまとめあげ、その観点で初めて明らかになるものを積み重ねていくことを目指しているのが、情報編纂である。

3. 研究会の目指すもの

情報編纂のコンセプトは、2006 年に提案を行い、2010 年に卒業した人工知能学会全国大会近未来チャレンジ「情報編纂の基盤技術」に選り、それも、2004 年 11 月に情報処理学会自然言語処理研究会・電子情報通信学会言語理解とコミュニケーション研究会の合同研究会で提案し、情報アクセス技術の研究促進を目的とした評価ワークショップ・シリーズ NTCIR^{*1} のパイロット・タスクとして、その後一部を正式タスクとして実施してきた MuST（動向情報の要約と可視化に関するワークショップ）と継承発展の関係にある^{*2}。

これまでも多くの研究者の方々が、コンセプトに共感し、研究素材に興味を持ち、様々なテーマで研究を展開して下さっている。2. 節で紹介した事例はその一端である。そのような中で、その核となるものを明らかにし、技術として展開していくための基盤を整理・整備することも必要と考え、人工知能学会の第 3 種研究会としての情報編纂研究会の発足となった。2009 年 11 月にキックオフ・シンポジウムを開催し [1] 活動を開始した。

研究会の目的として、以下の 4 点を挙げている [6]。

- 情報編纂技術の入門書（既存技術との関係や差異の明確化や、事例の紹介を通じて技術のコンセプトを明らかにし、多くの研究者にその分野に取り組む動機づけを与える方向性を持つもの）の執筆
- 情報編纂技術の展開を支援するインフラ（ツール群、プラットフォーム、コーパス）の構築
- 情報編纂技術の評価についての枠組みの提案
- 情報編纂技術をアピールするパイロットシステムの構築

つまり、情報編纂のコンセプトを明らかにし、様々な研究・技術を整理して情報編纂という視点での融合を図り、そのことの副作用として教科書が執筆される。その研究を進めるためのインフラであるツール群、プラットフォーム、コーパスについて、既存のものを整理するとともに必要なものを構築する。技術を評価する枠組みを定義する。そして、そのコンセプトを具現したシステムを例示することで応用と展開の方向を示そうと考えている。

現在の活動としては、まず、情報編纂に繋がる研究分野に関するチュートリアルや招待講演、情報編纂に関わる研究報告を中心としたシンポジウム（イベントとしての研究会）を開催している。これらを通じて情報編纂のコンセプトを明らかにしていく。以下にこれまでのチュートリアルと招待講演の一覧を示すが、ここにも情報編纂の関心が表われている。これらの内容や研究報告については、ホームページ^{*3}に詳しい。

- 統計的テキスト処理と情報編纂 金 明哲（同志社大学）
- チャンス発見において探索とは何か 〜Latent Dynamics へのシステムズ・アプローチ〜 大澤 幸生（東京大学）
- 情報編纂のためのテキスト処理技術 森 辰則（横浜国立大学）

^{*1} <http://research.nii.ac.jp/ntcir/index-ja.html>

^{*2} これらの活動についてはホームページ

<http://must.c.u-tokyo.ac.jp>,

<http://taurus.c.u-tokyo.ac.jp/InformationCompilation> を参照いただきたい。

^{*3} <http://taurus.c.u-tokyo.ac.jp/SigIC>

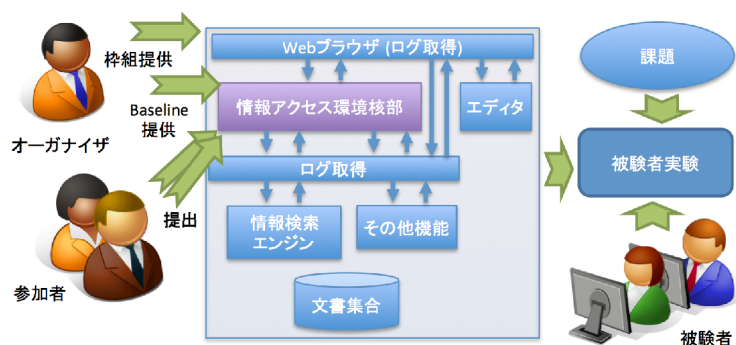


図 3: VisEx におけるシステム構成

- 情報編纂のための可視化技術 高間 康史 (首都大学東京)
- デジタルストーリーテリングと情報表現 角 薫 (一橋大学)
- 情報編纂と創造的テキストマイニングの融合による情報アクセス・分析環境 砂山 渡 (広島市立大学)

評価への提案やインフラの構築については、4. 節で述べる VisEx と並行して進めている。シンポジウムは、その報告やそれへのご意見を頂く議論の場としても活用していこうと考えている。多くの方々の研究活動と有機的に関連する研究会としていきたい。

4. 情報編纂評価への取り組み

VisEx (Interactive Visual Exploration Task) は対話的・探索的な情報アクセス環境の評価を行うことを目的として、NTCIR-9 においてパイロット・タスクとして実施されている^{*4}。評価ワークショップとは、定められたタスクを扱うシステムを参加者が持ち寄り、それらを共通の指標で評価したり、評価を通じてシステムの技術や評価方法についての議論を進めようという取り組みである。VisEx では、情報編纂というキーワードは前面に出していないが、情報アクセス環境の評価という目的、その中で視覚情報の利用を重視している点で、研究会の方向性と深く関連する。これを通じて、情報編纂技術の評価の枠組みについての議論を深め、そこに参加するシステムの構築を通じて、技術をアピールするパイロットシステムの構築も進めていきたいと考えている。

対話的探索的な情報アクセスの評価は難しい。2. 節で述べたように、その過程が以下の特徴を持つためである。

- その中で利用者の関心や求めるものが明確になっていたり変化したりするうえに、新しい知識が得られていくために、それらに伴った方向転換が生じたり、試行錯誤による一時的な逸脱とそこからの復帰が含まれたりすることも多く、直線的な行為ではない。
- キーワード等の条件の指定による文書の取捨選択だけでなく、サロゲートや文書の閲覧 (スキミングやブラウジング) 等、様々な手段 (視覚オブジェクトへの操作) が適宜組み合わせられて用いられ、均一な過程でない。
- 情報の理解や分析、傾向の把握や比較、集約と統合等、いわゆる情報検索とは違った、知識の組織化のための創造的な行為が大きな部分を占める。

このような複雑さのために、情報アクセス過程の結果として蓄積的に得られる情報の量や質で環境を評価しようという総合的な観点と、情報アクセス環境やインターフェースが持つべき機能や特徴をモデル化して整理し、それらについて評価しようという分析的な観点とが生まれてくる。その間の橋渡しは難しく、どちらかの観点に絞らないと評価が行いにくい。それぞれの観点に立ったとしても考慮すべきことは多い。総合的な観点は、一見精度再現率に基づく対話的でない情報検索の評価と似ているように見えるが、人間である利用者の存在、時間的要因、集約的包括的情報の扱い等を考える必要がある [3]。分析的観点にしても、そこで用いられる視覚オブジェクトがプレゼンテーションと対話的なインターフェースというふたつの側面を持つことから、少なくとも表現能力と対話機構の両方を意識したモデル化が必要となる [4]。

VisEx は、前者の総合的な観点での評価を後者の分析的観点での評価に結び付けることを最終的な目的とした取り組みであり、今期 (2011 年 12 月まで) は、その第一歩として被験者実験に基づく総合的な評価を実施する。実験は TREC^{*5} の interactive track [3] を参考にしたもので、共通の課題を被験者に与え、評価するシステムを用いて、その解決に繋がる情報をできるだけ多く制限時間内に集めるように指示するものである。命題にほぼ相当する小さい単位で情報を数え、その時間的蓄積を測定する。課題実施後には、利用者 (被験者) にアンケート調査を行い、システムの使用感や達成感のような主観的評価も収集していく。

できるだけ変数を減らし、かつ、情報アクセス過程全体を評価するために、図 3 のようなシステム構成を前提とし、情報アクセス環境核部を参加者に提出して頂く。特徴的であるのは、

- 背後にある情報検索エンジンと知識の編集や記録を行うエディタ部分を共通とすること
- 利用者の行動と検索エンジンの動作の両方の記録をとること

で、その仕組みについてはオーガナイザが提供する。このような構成とすることにより情報アクセス環境での利用者の行動が均質に取得できる。それが参加システムの総合的な評価になると同時に、分析的な評価のためのモデル構築に有効な指針を与えてくれることを期待している。更にこれらと、利用者 (被験者) の主観的評価とを関連づけていくことを目指す。

実験に用いる課題は図 4 の 2 種類で、共に新聞記事集合を文書集合として実施する。視覚オブジェクトの活用が期待で

*4 <http://must.c.u-tokyo.ac.jp/visex>

*5 <http://trec.nist.gov>

イベント収集課題 ある条件を満たす出来事群に関する情報を所定時間内にできるだけ多く快適に獲得する。

- 世界各地でなされた恐竜の化石の発掘について、その日時や場所、また発掘された恐竜の種類や生息時期について調べてください。
- アジアでの航空機墜落事故について、その発生日時、場所、航空会社名および死傷者数を調べてください。
- 携帯電話事業で国際的な提携を行った会社やその提携内容について調べてください。
- アジアで発生した地震について、その日時、場所、地震の規模および被害状況を調べてください。
- 世界各国で行われている核実験について日時や場所を含め調べてください。

トレンド要約課題 時系列統計情報が関連する社会や経済の状況について、その変化（動向）と変化の原因や影響を適切かつ快適に要約する。

- 固定電話、PHS、携帯電話等、通信関係機器の加入状況を調べてください。
- 内閣支持率や不支持率の変化を調べてください。
- ビール及び発泡酒のそれぞれの出荷量とそれらの比率、メーカー毎の状況について調べてください。
- 原油価格とガソリンの全国平均店頭価格の動向を調べてください。
- 失業者数や失業率等、雇用状況の変化を調べてください。

図 4: VisEx のふたつの課題

き、複数の文書に分かれた出来事や事態に関する情報をまとめる等、集約と統合という側面を含むものとなっている。新聞記事は 1998-2001 年のものを利用し、10 年前の記事ということで、予想しなかった事実が見つかったり、新しい解釈が生まれる等、被験者に何らかの発見があり、それによる振る舞いの変化がおこることも期待している。また、web 文書と異なり文書間にリンクは存在しないので、ブラウジングを可能とするような仕組みは情報アクセス環境核部が提供する必要がある。

VisEx は本年 7 月に被験者実験を行い、そこから得られた知見を 12 月のワークショップで報告する。あわせて、既に実施済みの予備実験の結果や環境の設計について、情報編纂研究会シンポジウム場を使って報告し、議論を進めていく予定である。

5. おわりに

情報編纂がどのような技術なのか、それに対して研究会としてどう取り組んでいるかについて述べた。様々な情報を編集・編纂し、わかりやすくするために多くの技術が関わっており、それらに対話的・探索的な情報アクセスの過程を支援するという観点で統合していく必要がある。そして、その評価については、インタラクションや人間の情報探索行動への知見に基づく検討が必要とされる。本稿をきっかけとして、より多くの方がこの技術に関心をもって頂き、議論に加わって頂ければ幸いである。

謝辞

情報編纂研究会は、松下光範氏（関西大学）、森辰則氏（横浜国立大学）、高間康史氏（首都大学東京）、阿部明典氏（NTT）が幹事・コアメンバとなって活動しています。VisEx は松下光範氏、上保秀夫氏（筑波大学）がオーガナイザを務めています。本稿で述べられている考えはこれらの皆様との活動の中で生まれてきたものです。加えて、本稿執筆についても貴重なご助言を頂きました。深く感謝します。本研究の一部は、基盤研究 (B)「視覚情報を利用した対話的情報アクセスのための情報編纂研究基盤の構築」として実施されており、科学研究費補助金を受けています。ご支援に感謝します。

参考文献

- [1] 阿部 明典: イベントだより 情報編纂研究会キックオフ. 人工知能学会学会誌, Vol. 25, No. 2, p. 301, 2010.
- [2] M.J. Bates: The Design of Browsing and Berrypicking Techniques for the Online Search Interface. in Online Information Review, Vol. 13, No. 5, pp. 407-424, 1989.
- [3] S.T. Dumais and N.J. Belkin: The TREC Interactive Tracks: Putting the User into Search. in E.M. Voorhees and D.K. Harman ed. TREC Experiment and Evaluation in Information Retrieval, pp. 123-152. The MIT Press, 2005.
- [4] C.M.D.S. Freitas, P.R.G. Luzzardi, et al.: On Evaluating Information Visualization Techniques. in Procs. of the Working Conference on Advanced Visual Interfaces, pp. 373-374, 2002.
- [5] 加藤 恒昭, 松下 光範: 時系列情報の抽出と可視化に基づく情報アクセスのためのマルチモーダルインタフェース -情報編纂の基盤技術に向けて-. 人工知能学会論文誌, Vol. 22, No. 5, pp. 553 - 562, 2007.
- [6] 加藤 恒昭: 研究会総覧 情報編纂研究会 (SIG-IC). 人工知能学会学会誌, Vol. 25, No. 4, p. 570, 2010.
- [7] 森 辰則, 藤岡 篤史, 村田 一郎: 動向情報編纂のためのテキストからの統計量表現の自動抽出. 人工知能学会論文誌, Vol. 23, No. 5, pp. 310-318, 2008.
- [8] 難波 英嗣: 動向情報の抽出と要約 -動向をまとめる-. 知能と情報 (日本知能情報ファジィ学会誌) Vol. 22, No. 5, pp. 549-555, 2010.
- [9] 高間 康史, 山田 隆志: 時空間的動向情報の探索的分析を支援するインタラクティブな情報可視化システム. 人工知能学会論文誌, Vol. 25, No. 1, pp. 58-67, 2010.