解説

動向情報の要約と可視化
－言葉と図で情報をまとめる－

加藤恒昭　東京大学 <kato@boz.c.u-tokyo.ac.jp>
松下光範　NTT コミュニケーション科学基礎研究所 <mat@cslab.kecl.ntt.co.jp>
神門典子　国立情報学研究所 <kando@nii.ac.jp>

■動向をまとめるということ■

WWW をはじめとして，さまざまなメディアによる膨大な情報が我々の周りに満ちあふれており，多種多様な情報から必要なものだけを取り出し，その内容を素早く理解したり概観したりすることがますます重要となっている．本稿では，あるトピックに関する一定の期間にわたる情報を動向としてまとめあげる技術について解説する．「動向」とは，広辞苑によれば「事態の情勢，または個人，集団などの行動の現況や将来の方向，なりゆき，うごき」とある．ここでは「将来の方向」は割愛して，いくつかの統計量の時系列データを基として，その変化を通時的にとらえつつ，それらを単に羅列するのではなく，総合的にまとめあげることで得られる概要を動向とする☆1．商品の売り上げ動向，政党支持に関する動向などがその典型である．動向情報は，さまざまな情報源からのある程度の期間にわたる情報を吟味して初めて得られるものであるから，それを自動的に生成することは利用者の関心を反映した膨大な情報の理解や概観のための技術の典型例であり，その必要性は大きい．

このような動向情報のまとめあげに必要な技術として，膨大なテキストの内容を簡潔にまとめることを目的とするテキスト要約や，多量の情報を視覚的に表現することでその理解とそれへのアクセスを支援する情報可視化が挙げられる．本稿では，これらの技術について簡単に説明した後，言語情報と視覚情報と，扱う情報の種類は違うものの，これらの技術が情報の迅速な理解や概観という同じ目的を持つことから，これらを融合したマルチモーダル要約が有益かつ必要であることを述べる．特に動向情報は，その表現においてもその素材となる情報においても数値情報・視覚情報等の非言語情報と言語情報との両方を適切に利用することが必要であり，要約と可視化，そしてその融合であるマルチモーダル要約の利用価値が高い．動向情報の要約と可視化に関する技術の具体例を述べた後，そのような研究を加速することを目的に展開されているワークショップを紹介し，最後に動向情報の要約と可視化の展開にふれる．

☆1 要約と可視化の出力として得られるものが動向情報である．「動向情報の要約と可視化」は動向情報を生成するための要約と可視化であり，処理の対象・入力となる多量の数値データや状況の記述が動向ではない．

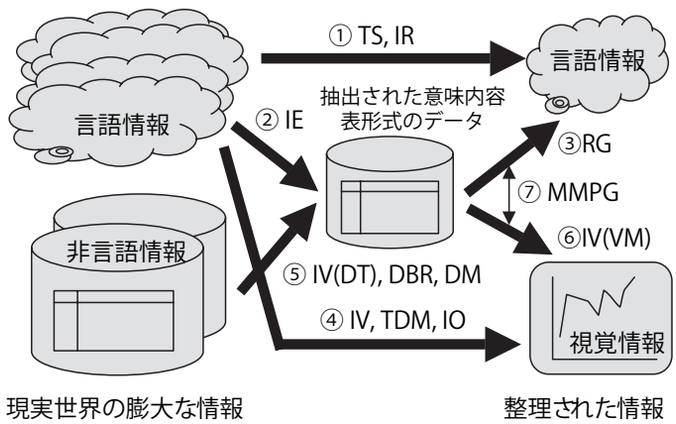
☆2 省略されているが重要な技術として，情報の意味的な注釈にかかわるセマンティック Web 技術，半構造化情報からの情報抽出に関するラッパー構築技術，自然言語質問文による情報アクセスを可能とする質問応答技術等が挙げられる．

■要約と可視化■

情報アクセス技術としての位置づけ

膨大な情報から利用者に必要なものだけを取り出し，その内容を素早く理解したり概観したりすることを可能とする技術，広い意味での情報アクセス技術の一部をまとめたものを**図 -1** に示す☆2．図の左側がさまざまなメディアによる膨大な情報の集まりであり，図の右側に示すのは，利用者の関心や興味に従って必要となるものだけに絞り込まれ，理解し概観しやすい形式で整理された情報である．この左側の膨大な情報から絞り込まれ整理された右側の情報を生成もしくは抽出し提示するための技術をまとめている．

図中矢印①のテキスト要約と情報検索は，膨大な言語情報からその要点となる情報を抽出し，それを言語情報として利用者に提示する技術である．情報検索が文書



TS	テキスト要約
IR	情報検索
IE	情報抽出
RG	レポート生成（テキスト生成）
IV	情報可視化
TDM	テキストデータマイニング
IO	インフォメーション・アウトライニング
IV(DT)	情報可視化におけるデータ変換
IV(VM)	情報可視化における視覚情報への写像
DBR	データベース検索
DM	データマイニング
MMPG	マルチメディアプレゼンテーション生成

■図 -1　さまざまな情報アクセス技術

を単位として利用者の関心に適合する情報を抜き出すのに対し，テキスト要約は，文やそれより小さい単位で情報を取り出し，再構成する．従来のテキスト要約では単一の文書を利用者の関心とは関係なく要約することが研究の主流であったが，現在は複数文書を対象とした研究が主流で，利用者の関心に注目して利用者が必要とする情報をまとめることにも関心が高まっている[4),5)]．テキスト要約は言語情報から言語情報への直接的な変換を主な手法とし，一般にその過程にテキストの意味理解の結果に相当する意味内容の表現を介さない．一方，矢印②に示す情報抽出はあらかじめ定義された意味内容の枠組みを埋めるかたちでテキストから必要な情報を取り出すという意味理解を行う☆3．どんな情報を抽出するかの定義があらかじめ必要でそれを受けて開発されるので，個々のシステムは分野依存で，任意の文書群を対象とできるテキスト要約とはその点でも対比される．ただ，最近は，抽出すべき情報の枠組み自体を文書群から自動的に抽出することを試みているものもあり，情報抽出におけるシステムの分野依存性は必ずしも必然ではなくなっている[6)]．また，矢印③として示された，表形式のデータや意味内容から自然言語のテキストを生成する技術は，テキスト生成，特にレポート生成と呼ばれるが，この技術を情報抽出と組み合わせることで，つまり，矢印②＋③という組合せで，矢印①のテキスト要約と同様に，膨大な言語情報からその要点となる情報を抽出しそれを言語情報として提示することが可能である．この場合は，意味内容の表現を介した要約が行われることになる．そのような情報抽出に基づいた要約も数は多くないがテキスト要約の研究として検討されている．

情報可視化はさまざまな側面を持っている．矢印④は，内容等に基づいた文書間の関係や文書に含まれる概念や語の関係を視覚化する研究である．このような文書に関する情報可視化は，文書集合から新しい事実やパタンを見つけ出す等の探索的データ分析を支援するテキストデータマイニングの一部として位置づけられることもあり，情報アクセスの効率的なインタフェースを構成するものとして，インフォーメーション・アウトライニングと呼ばれることもある．一方，組織化された非言語情報，たとえば膨大な数値データや 5W1H に分解された事実の列挙（地震発生や台風上陸の情報等）の情報可視化の研究も盛んである☆4．そのような可視化は膨大な非言語情報からその意味づけが明らかな表形式のデータを介して視覚情報へ至る処理となる．その過程は矢印⑤＋⑥ で示されるが，まず必要なデータを取り出しその意味付けを明らかにするデータ変換（矢印⑤）があり，その後の視覚情報への写像（矢印⑥）が続く．データ変換は，膨大な非言語情報から必要かつ価値ある情報を取り出すという点でデータベース検索やデータマイニングともとらえることができる．なお，図には示されていないが，このような可視化では利用者が視覚情報を対話的に操作して利用する視覚情報の変換が続く．これらには，シミュレーション実験の結果等の理解を支援するための科学的可視化と不動産データベースのような数値データへの対話的なアクセスを支援するものとがある[1)]．

☆3 情報抽出における意味理解は一般に浅い処理，簡単な処理，によって行われ，言語理解における深い理解とは区別されることが多い．

☆4 画像情報，映像情報，音声情報も重要な非言語情報であり，それらに関する技術の研究も盛んであるが本稿では割愛する．

このように見ると，テキスト要約と情報可視化は，出力する情報が言語情報と視覚情報ということで大きく異なることを別にすれば，きわめて近い目的を持っており，お互いの過程を対応づけられることが分かる．また，現実世界においては，情報が言語情報のみ視覚情報のみで提供されることは稀で，グラフを伴ったレポートや適当な注釈がついた図面など，それらを協調させたかたちで利用されるのが一般的である．そのようなマルチメディアプレゼンテーションを利用者の関心や意図を考慮して意味内容から自動生成しようというマルチメディアプレゼンテーション生成(矢印⑦)に関する研究も続けられている．そこでは意味内容から視覚情報を生成する視覚情報への写像(矢印⑥)と意味内容からテキストを生成するレポート生成(矢印③)の技術を有機的に組み合わせ，協調させることが必要となる．

要約と可視化の融合の意味

言語処理技術をはじめとするさまざまな技術に対する現在の期待は，目の前にある膨大な情報をどう処理するかにあり，実世界に存在する膨大な数値データや組織化されていないテキストを扱える適用範囲の広い頑強な技術が必要とされている．その点で，求められるのは整った意味内容を前提とするレポート生成（テキスト生成）やマルチメディアプレゼンテーション生成ではなく，任意のテキストを対象としたテキスト要約であり，膨大な文書集合や数値データを扱う情報可視化である．言語情報と非言語情報を協調的に扱うことの重要さを考えると，テキスト生成に対するテキスト要約の関係をマルチメディアプレゼンテーション生成技術に対して持つマルチモーダル要約技術[☆5]が望まれる．言い換えると，整った意味内容を前提とするマルチメディアプレゼンテーション生成技術も，実世界のさまざまな情報を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へと発展させる必要がある．

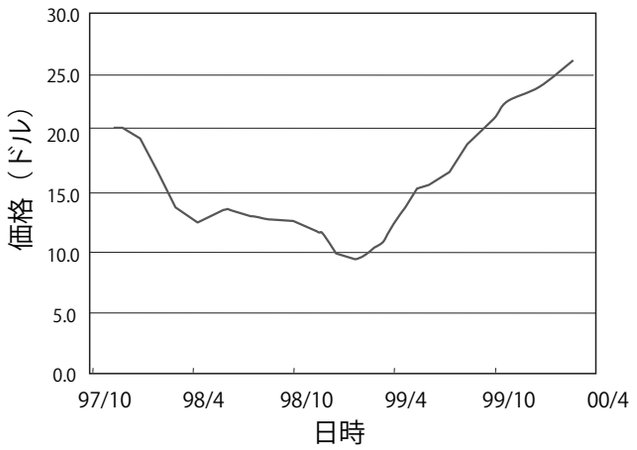
このマルチモーダル要約技術は，言語情報と非言語情報を扱うということでテキスト要約と情報可視化の融合となろうが，前節で述べたようにこの２つの技術はきわめて近い目的を持っており，お互いの過程を対応づけられるので，その融合は単なる足し合わせにはとどまらない．たとえば，図 -1 の矢印②＋③で実現される要約と矢印⑤＋⑥で実現される情報可視化とは同じ形式の中間結果を持つので，②＋⑥，⑤＋③という流れにより言語情報と非言語情報との相互変換が実現できることに加えて，②，⑤で得られた中間結果を総合的に処理して適切な補完等を行い，矢印③，⑥の入力として割り当てること等も可能である．矢印①に示される言語情報の直接的な変換による要約と矢印②＋③で示される表形式の意味内容を介在させる要約とは必ずしも対立的なものではなく中間的な手法が考えられるので，それらの要約と情報可視化との融合はさらに異なる形をとる可能性もある．加えて，矢印②と矢印⑥の組合せにより言語情報から視覚情報が生成できるが，これは矢印④と同じ入出力を持つので，その間でも有機的な融合が考えられる．

情報理解と情報アクセスの支援

テキスト要約と情報可視化の目的は情報の理解と情報へのアクセスを支援することである．テキスト要約では原文の代わりに用いられてそれだけで原文の内容を理解するための報知的要約と，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られる指示的要約が区別される．前者が情報の理解を，後者が情報へのアクセスを支援すると考えられる．ただし，一般に指示的要約よりも報知的要約の方が難しいので報知的要約が研究の関心であることが多いうえ，指示的要約をどのように利用して原文に効率的にアクセスするかはテキスト要約研究の外側にある．一方，情報可視化はさまざまな側面を持ち，情報理解のためのプレゼンテーション生成や情報へのアクセスを支援する視覚的インタフェースの構築が含まれる．前述のように情報可視化の過程には視覚情報の変換が含まれ，情報理解のためのプレゼンテーションを対話的に操作して情報アクセスのインタフェースとして用いることを可能とする．また，図 -1 の矢印④のインフォメーション・アウトライニングは情報アクセスを支援するが，そこでは指示的要約が文書の特徴として利用されることもあり，その意味ではテキスト要約が情報可視化の要素技術として振る舞うこともある．

情報の迅速な理解と概観といった場合，それに必要な情報の深さや細かさ，理解や概観の次に何をするかという目的は利用者によって異なるであろうから，情報理解のためのプレゼンテーションは，対話的な利用を前提とすべきであり，その基となった情報へのアクセスの支援と切り離すべきものではない．したがって，マルチモーダル要約の研究も情報理解と情報アクセスという２つに対する支援を視野に入れるべきであろう．この点ではテキスト要約よりも情報可視化研究のアプローチに近い立場をとるべきと考える．マルチモーダル要約は対話を利用した情報アクセス支援を考慮に入れる点でも，マルチメディアプレゼンテーション生成やテキスト要約とは大きく異なり，それらを拡張したものとなる．

☆5 マルチメディア要約は映像情報の抜粋を指すことが多いので，それと区別するためにマルチモーダル要約という用語を使う．



■図 -2　ドバイ原油価格の変化

■動向情報を扱う■

動向情報に何を求めるか

図 -2 は 1998 年から 1999 年にかけてのドバイ原油価格の推移を示している．以下は言語情報によるその要約である．

原油価格は 98 年には下げ続け，98 年末には 1 バーレル＝ 10 ドル台に落ち込み，99 年 2 月には一時 10 ドルを割り込んでいた．その後，4 月から 5 月にかけて 50%も上昇し，15 ドル前後となった．上昇は止まらず，8 月後半に 1 バーレル＝ 20 ドル 20 セントと 20 ドルの大台に乗り，98 年 2 月以来の高値を更新，9 月後半には 21 ～ 22 ドルとなった．

すでに述べたように，本稿での動向とは，いくつかの統計量の時系列データを基として，その変化を通時的にとらえつつ，それらを単に羅列するのではなく，総合的にまとめあげることで得られる概要である．ドバイ原油価格が基となる統計量の時系列データである場合，原油価格の「動向」としてはどのようなものが適切だろうか．図 -2 のグラフがすでに動向を表していると考える人もいるだろう．数値の列挙である表形式のデータに比べて，視覚情報はその全容を直観的に把握しやすく，「総合的なまとめあげ」の性格を持っている．一方で，上に示したテキストには変化の節目節目への言及がありそれがあって初めて「概要」であり，どこに着目すべきかが明らかでないグラフよりも動向としてより優れているという評価もあり得る．このように視覚情報と言語情報の性格，得手不得手を考えると，動向の適切な表現にはそれらの協調が必要であることが分かる．具体的には，テキストによって注釈されたグラフや，テキストを中心にそれを補強するグラフ等が有効であろう．

動向情報の表現は言語情報，視覚情報のいずれについてもさまざまな広がりを持っている．言語情報では，上例のような具体的な値への言及や変化の傾向に関する記述にとどまらず，その原因や評価，さらにはその影響等がまとめられるべきである．原油価格の場合，それはとにかくモノの価格であるので 50%，100%と上昇すれば大きな影響があるだろうと推測されるが，一般にはある数値の変化がどのような意味を持つかは自明でない．視覚情報も単純な折線グラフとは限らず，原油価格とガソリン小売価格との比較や，製品シェアや政党支持率で複数の会社や政党の間での比較が必要となる場合はそれに応じたグラフ形式を用いる必要がある．土地の価格動向では空間的な情報も表現しなければならないし，地震発生や台風上陸の傾向を動向として表現する場合は，その広がりはさらに大きい．パソコン業界の動向，通信と放送の融合に関する動向等，複数の統計量を組み合わせてそれらの相互関係とともに説明すべきものもあり，その場合は用いられる言語情報も視覚情報もより複雑なものとなろう．

また，動向はそれ自体で情報であるが，より詳細な情報へアクセスするインタフェースとしての役割も持ち得る．そのため，単なる表現，プレゼンテーションにとどまらず，利用者の関心の違いや変化に応じてさまざまな詳細度で対話的に情報を眺められることも必要である．視覚情報であればズーミング操作や異なるグラフ形式への変更を許すことでそのような対話が可能となるし，言語情報の場合，何を内容に含めるかについて観点の異なる要約を提供したり，要約の基となった原情報へのアクセスを可能とすることが必要である．この点で，報知的要約と指示的要約が縫い目なく繋がることが望ましいし，情報理解とあわせて情報アクセス支援の側面が重要となる．

動向情報をどこに求めるか

動向情報はその基となるデータも言語情報と非言語情報にまたがり，しかも複数の情報源に分散している．たとえば，図 -2 のようなグラフを描くためには白書等にある数値情報を用いるのが容易であるが，数値情報からだけでは，どこがその節目であるか，変化や値に対してどのような解釈・評価をすべきかを得ることは容易ではない．新聞記事等はそのような節目に現れ，その原因や影響についての記述を与えてくれる．動向情報を手掛かりとして詳細情報や関連情報にアクセスしたい場合も，新聞記事のような言語情報から得られる情報は重要で

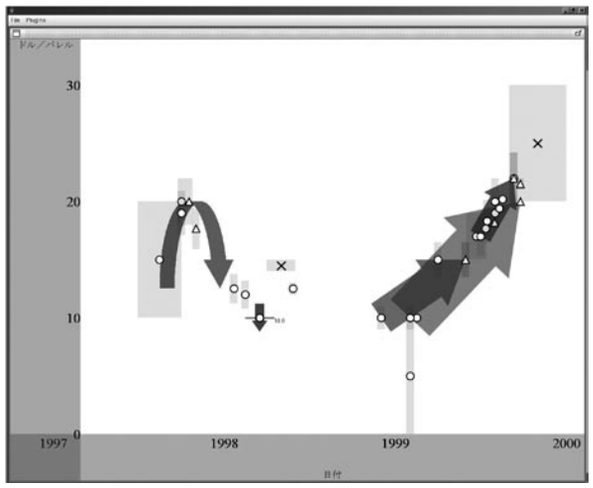
ある，

新聞記事のような文書の集まりから動向情報をまとめることはテキスト要約と情報抽出の技術にとっても興味深い課題である．動向はある程度の期間にわたる情報を吟味して初めて得られるものであるし，それぞれの文書は必ずしもその動向に関する内容だけを含んでいるわけでないので，複数文書を対象とした利用者の関心を考慮したテキスト要約が必要である．情報抽出技術としては，特定の統計量に特化したシステムではそもそも意味がないので，動向情報全般を扱うための汎用化が必要である．個別の出来事の情報ではなく，ある統計量の一連の値を抽出するという点も新しい．

さらに，動向情報の扱いにおいては，情報抽出技術とテキスト要約技術にさまざまな相互作用の可能性があり，動向情報の要約と可視化が単なる既存技術の足し算ではないことを示唆している．たとえば，ドバイ原油価格に関する記事では，「2 月には 10 ドルだったので」「今年 2 月には一時 10 ドルを割り込んでいたが」「2 月の 1 バーレル＝ 10 ドルから倍以上に高騰し」のように 99 年 2 月の底値が繰り返し言及されている．ここでこの底値を取り出すことは情報抽出技術であるが，この時期のこの底値がひとつの節目でありその値が動向情報の一部として必要であることを判断するのはテキスト要約技術であろう．また，ドバイの原油価格はその後上昇を続け，次々と異なる価格が報告され続けるが，それらの数値は「急騰している」「原油続騰」「急騰を続け」「上昇が止まらない」等々の表現とともに現れる．このような状況からこの時期に一定の変化，つまり急騰が続いていることを認識し，その期間中のそれぞれの時点の価格はあまり重要でないと判断することにもテキスト要約技術が利用されることになろう．

■動向情報を扱う技術■

動向情報を扱っている技術の具体例をいくつか見てみることとする☆6．**図 -3** に示すのは動向情報のための情報抽出と情報可視化に関する 1 つの提案である．グラフには数種類の点，矩形，形状の異なるいくつかの矢印記号が示されている．点は言語情報から抽出された統計量データを示している．点の種類は，それが文書中に直接表現されたものか，他の時期との比較等の記述を利用してあるいは分野知識を利用して推論や演算によって間接的に求められたものか，文書中で予測あるいは見込みとして述べられたものかに対応している．それぞれの表現の例を以下に示す．



■図 -3　視覚情報による動向の表現例

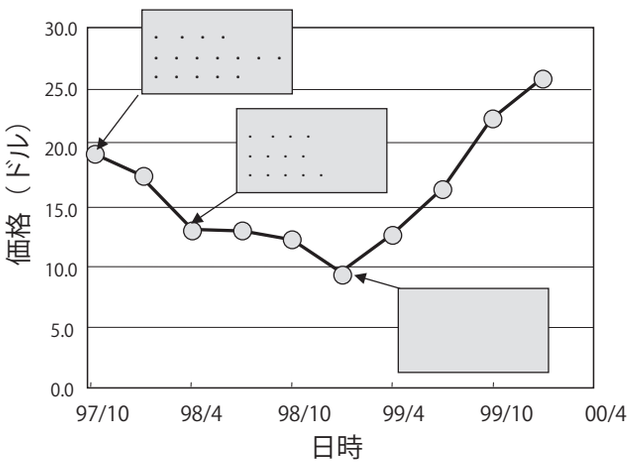
直接表現「先週末，1 バレル＝ 20 ドル 20 セントと 20 ドルの大台に乗り」

比較表現「2 月の原油価格は昨年 10 月より約 40％の下落になっている」

予測表現「現状の 1 バレル＝ 20 ドルを中心とした水準で堅調に推移するだろう」

図中の矩形は「22 ドル台」「21 ～ 22 ドル」「15 ドル前後」等の概然表現から得られたものである．統計量の値だけでなく「今年前半」「夏頃」のように時間表現が蓋然的である場合もある．そして，矢印記号は「97 年 10 月をピークに下落している」「10 ドルを下回った」「急騰を続けている」等の定性的な記述から得られた情報を表現している．これらの表現は基本パタンとして定義されたグラフ概形のいずれかに対応づけられ，そのパラメータ，ピークの時期やその際の値，何ドルを下回ったか

☆6 ここで述べるシステムや技術の多くは 2006 年 3 月に行われた言語処理学会第 12 回年次大会併設ワークショップ「言語処理と情報可視化の接点」で発表されたものである．その詳細についてはワークショップ予稿集を参照いただきたい．



■図 -4　グラフへの言語情報による注釈

等が具体化され，グラフに貼付される．このように，一般の情報抽出が対象とする個別の事実にとどまらないさまざまな情報が抽出され，動向としてグラフにまとめあげられている．特に定性的な記述をグラフ概形に対応づけて表示することにより言語情報としてまとめあげられた概要を数値データと同じプレゼンテーションに融合することを可能としており，その結果，まさに動向情報がこの 1 枚のグラフに表現されている．

非言語情報と言語情報の融合は，抽出された数値データを描画したグラフを言語情報によって注釈することによっても可能である．**図 -4** にそのような注釈付きのグラフの様子を示す．値の変化の大きな点，減少から増加に転じる極小点，描かれている部分の両端等，利用者が関心を持ちそうな点にあらかじめ注釈を付与しておくことができる．注釈の内容としては，記事中に存在するその時点の変化の定性的な記述，原因や影響に関する記述等が考えられる．具体例として，内閣支持率の動向で，どのような事件がそれに影響を与えたかを示唆するためにその時点の支持率と関係が強い出来事を注釈することや，株価のように豊富な情報がある場合に，グラフ中の特定の時期に対応する複数の記事の要約結果を表示することとし，その要約率をグラフの粒度と一致させることで，グラフによって描かれる動向と言語情報の要約として得られる動向とを協調させる等が提案されている．

図 -3, 4 に示したグラフは，グラフ上の点を指定するとその情報を抽出した新聞記事が表示される等の仕組みを加えることで，情報アクセスのための対話的なインタフェースの役割を果たすこととなる．時系列データだけでなく，たとえば，地震の発生地域を示した地図情報は，それらの地震についての震度やマグニチュードの情報や，さらにはその発生を報じた記事，その影響を論じた記事等にアクセスするためのインタフェースとなり得る．地震情報の場合，どこで発生したかという地理的な情報といつ発生したかという時間的な情報をどのようにまとめるか，どのように視覚的に表現するかも興味深い．

これまでの事例は，原情報として言語情報を中心に扱っていたが，数値情報を対象としてそこから言語情報を生成し，グラフとそれを説明するテキストというかたちで視覚情報と言語情報とを協調させることも可能である．株価やガソリン価格を対象に，その数値情報から描かれるグラフを最少自乗法やファジィ集合の考え方を用いて分割し，それぞれの部分を特徴付け，それらを言語的に表現して，変化を説明するテキストを合成することが試みられている．

これらの研究の多くはまだ提案の段階であり，その実装や評価が十分に行われていないものもあるが，言語情報と非言語情報の協調の可能性，それによる動向情報の要約と可視化，それを通じた情報アクセスの大きな可能性を感じさせる．

■動向情報の要約と可視化に関するワークショップ：MuST ■

MuST の枠組み

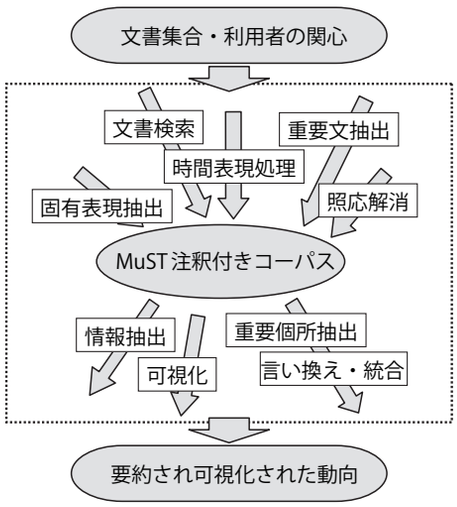
筆者らは，これまで述べたようなマルチモーダル要約，特に動向情報の要約と可視化の重要性と広がりに注目し，それに関する技術について，協調的かつ競争的に研究を進めていくための「MuST：動向情報の要約と可視化に関するワークショップ」（以下，MuST）を提案し，運営している [2) ☆7]．具体的には，共通の素材，研究用データセットを用いて緩い意味で共通の課題に取り組むことによって，議論と研究の活性化，コミュニティの形成や研究領域の認知度の向上，ツールやコーパス類の蓄積等を目指している．オーガナイザは研究用データセットの提供に加えて，メーリングリストや報告会等の議論の場の提供，研究成果発表の場の確保を行っている．参加者には提供されたデータセットを用いた研究の成果や経過について指定された機会に発表することを求めている．

MuST の研究用データセットは，ワークショップの求心力となり，動向情報の要約と可視化に関する研究を加速させることを目的に設計された．それは，研究の素材となる文書セットに注釈付けを行ったコーパスと出力の参考となる要約テキストやグラフ等からなっている．

☆7　http://must.c.u-tokyo.ac.jp

```
<unit stat="ドバイ原油価格 ">
また,
<name part="head">原油価格（ドバイ原油）</name>
も,
<date gra="月 " abs="199710">昨年 10 月ごろ </date>
<rft id="980214080 _ 1"><name part="foot">1 バレル=
  </name></rft>
<val>約 20 ドル </val>
をつけたのを <rel type="ord">ピーク </rel>
に下落が続き,
<date gra="旬 " abs="19980121">今年 1 月下旬 </date>
には
<pro ref="1 バレル " id="980214080 _ 1">同 </pro>
<val>約 12 ドル 50 セント </val>
まで落ち込んだ
</unit>
```

■図 -5　注釈の例



■図 -6　ハブとしての注釈付きコーパス

注釈付きコーパスは，ガソリン価格，パソコン出荷状況，ビール業界，台風，地震等，27 のトピックについて，新聞記事 2 年分から選択した関連記事からなっており，それぞれの記事では，そのトピックの動向に関連するであろう統計量に関する記述を取り出して意味的な注釈を加えている．その一部を**図 -5** に示す．

研究の素材，処理の入力となる文書セットにとどまらず，注釈付きコーパスを含めている点が重要である．この注釈付きコーパスは，あるコンポーネントの出力であり別のコンポーネントの入力となる処理の中間結果に相当し，動向情報の要約と可視化に関する研究の枠組みにおいてハブの役割を果たす．**図 -6** にその位置づけを示す．図では動向情報の要約と可視化に関する要素技術，処理を示しているが，その前半を構成する処理として以下があげられる．

- 必要な情報を含んだ文書を探し出し，それらの文書から重要部分を文単位で抽出する文書検索と重要文抽出
- 重要文を構成する要素についてそれが統計量名であるか日付であるか数値表現であるか等の意味付けを明らかにする固有表現抽出
- 「先月」「昨年同期」等の相対的時間表現についてその絶対表現の算出する時間表現処理
- 代名詞等についての照応解消処理

図 -5 の例から分かるように MuST コーパスの注釈付けはこれらの処理の結果に相当する．したがって，これらの研究に関心を持つ研究者にとっては，正解データあるいは学習用データとしての利用が可能である．そして後半の処理として，情報抽出や可視化，文より細かい単位での重要個所抽出と言い換えおよび統合による要約の生成があるが，これらに関心を持つ研究者にとっては，注釈の情報を利用することで，前半の技術開発に要する労力を節約して，自分たちの関心のある部分に直接取り組むことが可能となる．たとえば，可視化の研究を自然言語処理技術にかかわることなく進めることができる．加えて，数値情報に関するさまざまな言語表現に関する分析等も注釈の情報を用いることで効率的に進めることができる．動向情報の要約と可視化は，さまざまな研究分野にまたがったさまざまな要素技術を必要とし，そのシステム構築は必ずしも容易ではないが，このデータセットを用いることで，研究者は各人の関心ある要素技術に取り組むことが可能となる．

研究用データセットが求心力となると述べたが，より広い視野に立った時に重要なことは，この注釈付きコーパスを通じて，今まで異なる分野に属すると考えられていた研究者たちの議論が可能になるという点である．そして，もちろん，同じ分野の研究者は，このデータセットを共通の素材とすることで一定の客観的評価が行えることになる．これらのことを通じた研究の加速と活性化がデータセット構築の目的であり，MuST の目指すところである．

MuST の現状と今後

MuST は，2004 年 11 月に最初の提案を行い，2005 年初めからメーリングリストの立ち上げと参加者募集を行い，研究用データセットの配布を開始した．2005 年度のデータセットは，20 トピックを対象として 355 記事を注釈づけたものであった．15 組織からの

参加があり，2006 年 3 月には第 1 回成果進捗報告会が実施され，活発な議論がなされた．その概要についても MuST Web サイトから見ていただくことができる．あわせて，2006 年 2 月に開催された電子情報通信学会 NLC シンポジウムにおける関連テーマのセッションや，3 月の言語処理学会年次大会で関連ワークショップ「言語処理と情報可視化の接点」でも多くの発表がなされた．「動向情報を扱う技術」の章で説明した動向情報を扱うための技術のほとんども MuST をきっかけとして研究されているものである．

今年度は研究用データセットを追加し，昨年度のものとあわせて 27 トピック 581 記事の注釈付きコーパスを作成した．規模が大きくなっただけでなく，注釈仕様も精緻化し，特に統計量に関する表現以外の原因や影響に関する記述の分類も加えられている．参加申し込みは 7 月より始めている．NII が主催する NTCIR ワークショップ☆8 のパイロットタスクということで，毎日新聞記事の利用も可能となっている．昨年度同様，来年 3 月頃に成果進捗報告会の実施を考えている．研究用データセットを用いたあらゆる研究，システム構築，要素技術の確立，データ分析等々，を歓迎し，緩い意味で共通の課題に取り組むことによる議論と研究の活性化を目的とする．今年度は，そのような自然発生的な協調に加え，多くの研究組織が昨年度の研究によりさまざまな蓄積を行っているので，それら蓄積された資源の中から，たとえば一部の処理を実現するツールや評価用のデータ等を共有し，ワークショップ内で活用する枠組みも考えていく．

■動向情報の要約と可視化の展開■

テキスト要約と情報可視化の技術を協調させ，現実世界の膨大な情報に立ち向かっていく技術として，マルチモーダル要約を紹介した．動向情報は情勢や状況の概要であり，それを多量かつさまざまな情報から生成する必要性は大きいが，合わせて，その素材となる情報においても，表現の方式においても，言語情報と，数値情報や視覚情報等の非言語情報とがかかわっており，マルチモーダル要約の活用が期待される．ここでは統計量の時系列データに関連するものに限定したが，「動向」はそれ以上の広がりを持っている．多くの評価，意見等から形造られる人気や評判等，いわゆる流行やトレンドもそこに含めることができる．これらの要約や可視化の必要性も大きいが，それらを扱おうとすると，ここで述べた技術のさらなる展開や異なる技術の導入が必要と思われる [3]．本稿が動向情報の要約と可視化の重要性とその研究的意味を伝えており，多くの方が関心を持ってくださることを期待する．

謝辞　本稿で紹介した MuST は，NTT と東京大学との産学連携共同研究，ならびに国立情報学研究所と東京大学との公募型共同研究によって支援されています．ご支援をここに感謝します．本稿の内容には MuST 参加者からいただいた示唆に基づく部分が数多くあります．貴重な議論に感謝いたします．

参考文献

1) Card, S. K., Mackinlay, J. D. and Sheiderman, B. : Information Visualization, Readings in Information Visualization Using Vision to Think, pp.1-34, Morgan Kaufman Publishers, Inc. (1999).
2) Kato, T., Matsushita, M. and Kando, N. : MuST: A Workshop on Multi-modal Summarization for Trend Information, Proc. NTCIR-5 Workshop Meeting, pp.556-563 (2005).
3) 加藤恒昭，松下光範：情報編纂（Information Compilation）の基盤技術，第 20 回人工知能学会全国大会 1D3-2 (2006).
4) Mani, I. : Automatic Summarization, John Benjamins Publishing Co. (2001)（奥村　学，難波英嗣，植田禎子訳：自動要約，共立出版）
5) 奥村　学，難波英嗣：テキスト自動要約，オーム社 (2005).
6) 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).

（平成 18 年 7 月 27 日受付）

☆8 http://research.nii.ac.jp/ntcir/index-ja.html