動向情報の情報編纂

– NTCIR MuST を通じて見えてきたもの –

Information Compilation for Trend Information

加藤 恒昭[*1] 松下 光範[*2] 神門 典子[*3]
Tsuneaki Kato Mitsunori Matsushita Noriko Kando

[*1]東京大学 The University of Tokyo
[*2]関西大学 Kansai University
[*3]国立情報学研究所 National Institute of Informatics

A framework of information compilation for trend information is proposed. It has been inspired and developed through involvement in MuST workshop, which was designed to encourage cooperative and competitive studies on multimodal summarization and visualization for trend information. In this framework, three kinds of information concerning to a statistic are extracted from collected documents and organized into an interactive visual representation with numerical data. That representation can be regarded as an overview and also as an entry to further information access. This paper presents the outline of the framework and its relationship to the research resources constructed in MuST.

1. はじめに

著者らは，MuST（動向情報の要約と可視化に関するワークショップ）をオーガナイザとして運営し，同時に参加者としても関わってきた [5]．動向情報は，利用者の蓋然的な要求に対する最初の回答であり，それに続く対話的な情報アクセスの入り口となる．またその素材として数値データ等の非言語情報を含む場合も多く，言語情報と非言語情報の協調的な活用が重要となる．これらの点で，動向情報は情報編纂研究の格好の素材である．実際には動向情報の要約と可視化を行う技術をより一般化する形で情報編纂のコンセプトが提案されたという経緯となっている [1]．

本稿では，MuST での経験をふまえて，そこで行われた研究を紹介しながら，動向情報の情報編纂を行う枠組みを提案する．そして，そのような枠組みでのシステム構築のためにMuST を通じて設計・開発した資源がどのように有益であるかを説明する．

本稿の構成は以下の通り，まず 2. で，MuST 参加者の研究を参照しつつ，動向という概念の整理をし，提案する枠組みを位置づける．3. で，動向情報の素材となる情報の特徴を分析し，枠組みの提案を行う．4. で，MuST を通じて設計・開発した資源を紹介し，それが提案した枠組みに基づいた研究に有益であることを示す．5. でその拡張と展開について考察し，最後に 6. で全体をまとめる．

2. 動向の２つの捉え方

動向という言葉が何を意味するかについては MuST 参加者の中にも揺れがあり，大きくふたつの捉え方がある．ひとつの理解は，利用者の関心に応える情報の全体像を概観させるという役割を持ち,「今年に入って原油とガソリンの価格はどう動いているのだろう」「06 年からゲーム機業界はどんな感じになったのか」「去年の台風はひどかったのか」等で示される利用者の関心に対する最初の回答となるものとしての動向である．そこでは，蓋然的であるとはいえ，利用者の関心の指向性に導かれた情報アクセスが想定されている．

もうひとつの理解は，与えられた文書集合全体を対象に，そこにどのような情報や話題が含まれているか，それらにどのような関係があるか等を提示したもの，文書集合全体の概要，もしくはそれが扱っている世界や社会の情勢を概観するものとしての動向である．これに基づき，統計量名を実世界の出来事や現象を特徴づけるキーワードと捉え文書中から抽出し，その関係を可視化する試みや，文書集合中で言及されている時系列情報を含むすべての関係情報を抽出して可視化する取り組みが行われている [8]．まずは全体の概観ということで，利用者の関心や目的からくる指向性はなく，上昇的でデータ（文書中の情報）指向の情報アクセスとなる．このような２つの方向の情報アクセスについて，その入り口としての最初の応答が共に動向という言葉で括れることは興味深い．

前者の捉え方での研究は．利用者の関心となっている動向について，その動向を表現する統計量（原油価格等）や出来事（台風の上陸等）を表現する言語情報，非言語情報に関連した要約と可視化の検討が中心となる．本稿で提案する動向情報の情報編纂も時系列統計量に言及するテキストからの情報抽出とその可視化についての研究に動機づけられたものである．なお，ここにはひとつのギャップがあって，利用者の関心はある分野の動向であって，特定の統計量ではない．従って，政治の動向が内閣支持率や政党支持率によって示されることを判断するステップが必要である．動向と統計量の関係については 5. でも論じる．

3. 動向情報編纂の枠組み

時系列統計量である原油価格について述べている以下の文章を考えてみる．

> また、原油価格（ドバイ原油）も、昨年１０月ごろ１バレル＝約２０ドルをつけたのをピークに下落が続き、今年１月下旬には同約１２ドル５０セントまで落ち込んだ。その後、イラク情勢の緊迫化で一時上昇したものの、現在はまた１２ドル前後で低迷、「原油はだぶつき気味」（同庁）だ。(毎日新聞 980214080)

ここに含まれる情報は，少なくとも以下の３つに分類することができる．

1. 統計量名，時点，値の３つ組で表現できるようないわゆる統計情報（**時点数値情報**）．

連絡先: kato@boz.c.u-tokyo.ac.jp

2. 「ピーク」「一時上昇した」「低迷」のような統計量の変化を述べる情報（**変化情報**）．
3. 「イラク情勢の緊迫化で」「原油はだぶつき気味」等の状況変化の理由や状況に対する評価（**状況情報**）．

時点数値情報は，白書等の非言語情報から容易に獲得できるものではあるが，それを文書から抽出するのは，それが新聞記事等で言及されたことに意味があるためである．上例の「ピーク」であった「約２０ドル」からもわかるように，現れるのは統計量の動向を要約する節目節目の情報である．また，一定の間隔で調査や発表が行われる一部の統計量（内閣支持率等）は別にして，その時点の統計量が発表されたことにも意味がある[*1]．従って，このような情報の抽出とその結果の可視化によって，扱っている文書集合が持つ関心に従った動向情報の提示が可能となる．

この様な背景で，時点数値情報をテキスト中から抽出することが MuST 参加者の主な関心のひとつとなった．時点数値情報の抽出はいわゆる情報抽出の枠組みによって行われるが，以下に示すような難しさと同時に利用可能な特徴を有しており，それらに着目した研究が進められている (例えば [7, 6, 11])．

- 統計量名の正式な名称は長めの複合語（「レギュラーガソリン全国平均小売価格」）で構成されることが多いが，文書中ではそれがそのまま現れるとは限らず，異なる表現（「平均価格」と「価格の平均」）や省略が頻繁に見られるため，その同定が困難である．
- 同一記事内で複数の統計量に言及されることが多いため，ある数値がどの統計量の値であるかの同定が困難である．
- 時間表現は文脈を受けて省略されることが多く，広い範囲の解析が必要となる．加えて，発表日等，雑音となる日付もあり，同定が難しい．
- 一定期間にわたってある統計量の一連の値を抽出するということから，とりうる値の範囲や値の連続性などを仮定するヒューリスティクスを用いて，精度の向上を図ることができる．
- 以前の値との差等，相対的な値の表現も含まれており，それらに着目することで，再現率の向上を図ることができる．

変化情報は，以下のような興味深い特徴を持つ．

- 「つけた」「下落が続き」「落ち込んだ」という用言によって表現されることが多く，言語表現の分節の単位に対応している．
- 「今年１月下旬には同約１２ドル５０セントまで落ち込んだ」に見られるように時点数値情報は変化の終点や起点であり，変化情報が中心で時点数値情報はそのパラメータと考えられる．
- 時点数値情報に較べ，「ピーク」「低迷」等，ある時間幅や，「１０ドルを切った」等，特徴的な時点に関しての要約になっている[*2]．
- 「ピーク」「上昇した」のように客観的で類型化できるものに加えて，「落ち込んだ」「低迷」のように主観的な評価が上乗せされたものがある．

*1　ただし，これはその時点での判断で後に振り返った時にその重要性が維持されている保証はない．またこの情報は記事掲載日から復元できないこともない．

*2　文書中の時点数値情報も「１０ドル台」「約１０ドル」のような概数は要約的な性格を持つといえる．

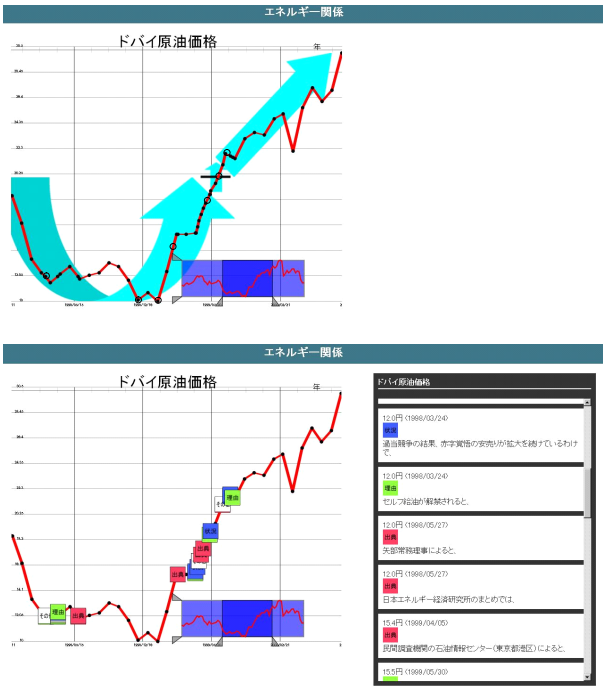

図 1: 動向情報の可視化

筆者らはその重要性に着目し，パタンに基づくその抽出を検討している [2]．

状況情報は変化情報によって示される状況に関連する言語的な記述となっている．これは主観情報抽出と同様の技術によって抽出が可能と考えられるが，MuST 参加者でこのテーマに取り組んだ例は見られていない．

これら３種類の情報を文書から抽出し，白書等から得られる統計量の数値データ（統計データ）と合わせて，時間軸にそれらを重ねて配置することで組織化し，この統計量に関する動向を可視化して概観を許す，同時に，この対話的なプレゼンテーションを詳細情報，例えば，記事全文へアクセスする際の入り口として用いるという枠組みが本稿での提案である．

このような枠組みでの表示例を図 1 に示す[*3]．図 1 上では．時点数値情報と変化情報が重ねられている．変化情報はその種類を表現する矢印アイコンによって示されている．ここには示されていないが，例えば，上向き矢印に「急騰」と注釈づけることにより，変化情報に含まれる主観的な評価を上乗せすることが可能である．

図 1 下では，状況情報が，理由，状況，出展等の種類で色分けされたアイコンでグラフと関係づけられている．ここでは，グラフ中のアイコンと右側に並んだスニペット（状況情報として抽出された部分）が関連づけられており，グラフ中で注目した状況情報のアイコンを指定することでその内容を知ることができるし，関心を持ったスニペットを指定することで，それがグラフ中のどの時点についてのものかを明らかにできる．更にそれらから，それを抽出した文，それが含まれる記事の見出し，記事全文を関係づけ，指定された場合にポップアップさせること等ができる．上においてもグラフ中の矢印アイコンや時点数値情報の点に，同様の関連づけが行える．記事の掲載時点ではなく，記事の内容が言及している時点に関連づけられる点が重要である．

*3　上例の文章で表現されたものとは異なる時期が表示されている．

<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>。<unit stat="ドバイ原油価格"><ins type="name"><name>ドバイ原油価格</name>は</ins><date gra="旬" abs="19980121">その</date>後、<del type="rsn">イラク情勢の緊迫化で</del>一時上昇したものの、<date gra="不明" abs="19980214">現在</date>はまた<val>１２ドル前後</val>で低迷、<del type="sit">「原油はだぶつき気味」（同庁）だ</del></unit>。

図 2: MuST データセットでの注釈

グラフの右下にあるのは，パンやズームのためのもので，これにより注目点を移動させたり，細部に注目したりすることができる．これらの画面を相互に行き来することも可能である．このような視覚的な関心の移動と，上で説明したアイコンの指定による文書情報への関心の移動を通じて，Shneiderman の情報可視化のマントラ “Overview first, zoom and filter, then details on demand”[9] が実現されることになる．

4. 動向情報編纂のための研究資源

前節で述べた情報編纂の枠組みを検討するために，MuST での共通の研究素材である *MuST* **データセット**と**変化表現コーパス**，多くの参加者が取り組んでいた問題を具体化した**評価課題**を用いることができる．

MuST **データセット**は，1998 年から 2001 年の毎日新聞記事を知識源と考え，そこから「ガソリン価格」「自動車生産」「通信機器」等，27 のトピックについて，関連する統計量や出来事（各トピックについて３つ前後，計 90 種類）に関する情報を含んだ新聞記事 702 記事を収集し，注釈づけを行ったものである[*4]．具体的な仕様は MuST Home Page[*5]や文献 [3] に詳しい．

トピック毎に収集した新聞記事集合は文書収集（情報検索）によって得られた言語情報に相当する．各記事への注釈は，要約における重要文の抽出と，情報抽出における固有表現抽出と時間表現解析を含む参照表現解析の結果に相当する．注釈の例を図 2 に示す．統計量や出来事に言及している文が unit 要素として抽出され，そこに含まれる統計量名（name 要素），時点（date 要素），統計量の値（val 要素）等が注釈づけられている．相対的な時間表現にはその絶対値が注釈される（abs 属性として yyyymmdd で表現）．これらは時点数値情報の３つ組を構成する要素となる．統計量の値や変化に直接言及していない部分は，del 要素とされるが，これが状況情報として抽出すべき部分に対応しており，type 属性によって，出展，原因（rsn），評価，状況（sit）等へと分類されている．

変化表現コーパスは，時点数値情報と変化情報の抽出結果とそれが抽出された言語表現を整理したものである．抽出される情報は，３つ組で表現されるある時点での統計量の値についての情報（type 0），値の変化に関する情報（type 1），変化の変化に関する情報（type 2）に整理されている．例えば，上例の「昨年１０月ごろ１バレル＝約２０ドルをつけたのをピークに」からはドバイ原油価格に関する以下の情報が抽出される．

type 0 時点=昨年１０月ごろ，時点との関係=一致，
値＝約２０ドル，値との関係＝一致
type 1 種別=下降
type 2 種別＝変化方向の転換

*4 MuST データセットは他にも幾つかの情報を含んでいる．
*5 http://must.c.u-tokyo.ac.jp

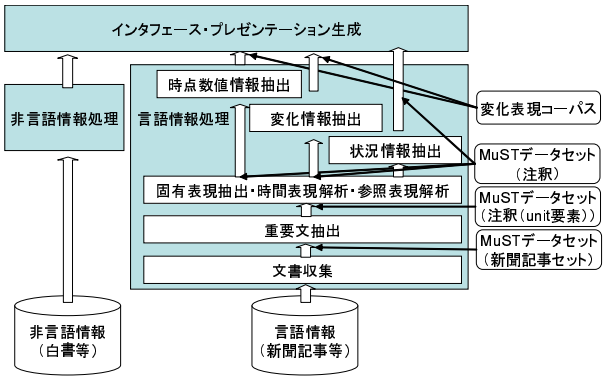

図 3: 構成要素と研究資源

また,「今年２月には一時１０ドルを割り込んでいたが、わずか半年で倍の値上がり（同 990817066）」という文章では，前半の「を割り込んだ」は，下降方向の変化があり，言及されていてる時点以前にその値となった時点があったと以下のように表現する．

type 0 時点＝今年２月，時点との関係=以前，
値＝１０ドル，値との関係＝一致
type 1 種別=下降

後半の「半年で倍の値上がり」は以下の様になる．

type 1 種別=上昇，期間=半年，比率=２倍

このように，値に関する様々な情報や，変化に関する定量的な相対情報が整理できるようになっており，時点数値情報と変化情報を網羅している．MuST コーパス中の 9 トピック 219 記事について，1,789 件が抽出されている．具体的な分析の方針とコーパスの仕様については文献 [2][4] に詳しい．

これら２つの研究資源と前節で述べた枠組みとの関係を図 3 に示す．多くの構成要素の出力に相当する情報をこれらの資源が提供していることがわかる．これらは，サブシステムの設計で参照したり，機械学習に基づく手法での学習データとして使うことが可能である．また，特定のサブシステムに研究を集中するために，それ以前の処理結果としてこれらを利用することもできる．

評価課題は動向情報の要約と可視化に関する要素技術について客観的定量的評価を行うものとして立案された．言語情報と数値情報の相互変換や対応付けを可能とするための以下の３課題である．

- T2N 課題（言語情報に関する情報抽出）
- ALN 課題（言語情報と数値情報のアラインメント）
- N2T 課題（数値情報の言語情報化）

T2N **課題**は，文書集合中で言及されている時系列統計情報を一定の形式で抽出する．その文書集合中で，ある統計量のどの時点のどの値が話題や関心となっているかを明らかにするもので，時点数値情報を抽出することに相当する．

ALN **課題**は文書集合中でなされている統計量やその変化への言及を取り出し，数値の列として表現されている時系列統計量（統計データ）の対応する部分に関連づける．メディア・アラインメントと呼べるもので，異なる情報源による異なるメディアの協同的利用を可能にする．この課題は，変化情報と状況情報の抽出と，その後の統計データとの組織化に関連する課題である．上例であれば「下落が続き」「一時上昇」「１２ドル前後で低迷」等が実際の統計データのどの部分に関する記述であるかを明らかにすることになる．

N2T **課題**は統計データの変化や概要を簡潔に表現する文章を生成する．得られる文章は数値情報の言語による要約である．この課題は，本稿で提案したものと双対をなす情報編纂のアプローチに関連する．

MuST において実際に実施されたのは T2N 課題のみで，評価も情報抽出で一般的な精度と再現率によって行われた．実施の結果は文献 [4] に詳しい．T2N 課題において動向情報であるという課題の特色を活かした設定や評価を検討することと，ALN 課題の具体化によって，前節で提案した枠組みの構成要素を客観評価可能な基礎技術として確立していくことが期待される．

5. 動向情報編纂の展開

本稿で提案した動向情報編纂の枠組みは更なる発展のための課題を残している．

時系列統計情報は多くの統計情報の基礎となるものであるが，統計情報は時間軸だけで表現されるとは限らない．政党支持率は時間だけでなく政党を，ある製品のシェアは企業をパラメータとして持つ．統計量のみが動向と関連するわけでもなく，地震や台風の動向を考えると，時間軸に加えて，震源地や上陸場所をパラメータとする出来事の表現が必要になる．このような場合に，多元的な可視化を対話的に操作し，時間軸，空間，企業の軸等々，様々な観点からその振る舞いを眺められる可視化が必要となる．そのときには注目点の移動や絞り込みにも工夫が必要となる．文献 [10] では，このような多面的で対話的な可視化を OLAP で用いられている概念を応用することでモデル化している．

ひとつの統計量の複数のパラメータだけでなく，複数の統計量を扱う必要がある動向も多い．原油価格はガソリン小売価格と較べることが必要となるし，政治動向においては，政党支持率と内閣支持率が並べて議論される．このような複数の統計量の可視化についてどのような配置や操作が可能かを検討していく必要がある．更にこの場合，それと組織化されるべき言語情報も，例えば「原油は上昇、なぜかガソリンは下落（990530053）」のように異なる統計量をまたがることになる．言い換えると，このようなコメントによって２つの統計量が関係づけられる．そのような関係づけをどのように示していくかが課題となる．

更に「生産のだぶつきやガソリンスタンドの過当競争が、特殊な状況をつくり出している。（同前）」のような統計量から離れた，動向そのものに関する状況情報の扱いも問題となる．この様な情報については MuST コーパスでも注釈付けをしておらず，整理の枠組みからの検討が必要である．

6. おわりに

本稿では，動向情報の情報編纂について，MuST での検討を通じて見えてきた枠組みを提案した．あわせて，その枠組みと関連する MuST 内で行われた研究を紹介し，その実現のために MuST を通じて構築した研究資源が活用できることを示した．加えて，その発展のための課題を述べた．今後は，このような枠組みを洗練して，要素技術を明確にすることで研究コミュニティを活性化すると共に，枠組みの発展と拡張を検討していきたい．

謝辞

MuST の運営とそれに関連する研究は，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の NTT 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いたします．MuST の活動全般と本稿の執筆は MuST に参加頂いた皆様によって可能となったものです．あらためて感謝いたします．

参考文献

[1] 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.

[2] 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.

[3] Kato, T., Matsushita, M. and Kando, N.: Expansion of Multimodal Summarization for Trend Information –Report on the First and Second Cycles of the MuST Workshop –. in Procs of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pp. 235–242, 2007.

[4] Kato, T. and Matsushita, M.: Overview of MuST at the NTCIR-7 Workshop – Challenges to Multi-modal Summarization for Trend Information –, in [8], pp. 475-488, 2008.

[5] 加藤 恒昭・松下 光範・神門 典子「動向情報の要約と可視化とその展開 – MuST（動向情報の要約と可視化に関するワークショップ）活動報告–」情報処理学会自然言語処理研究会, 2009-NL-190-12, 2009.

[6] Mori, T. and Miyazaki, R.: A Simple Baseline Method for NTCIR-7 MuST T2N Task – Yokohama National University at NTCIR-7 MuST T2N –, in [8], pp. 502-508, 2008.

[7] Nanba, H.: Extraction of Trend Information from Newspaper Articles: Hiroshima City University at NTCIR-7 MuST, in [8], pp. 489-493, 2008.

[8] National Institute of Informatics: Procs of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies, 2008.

[9] Shuneiderman, B. and Plaisant, C.: Designing the User Interface (fifth edition). Addison-Wesley, 2009.

[10] Takama, Y. and Yamada, T.: Interactive Information Visualization of Trend Information, in [8], pp. 528-533, 2008.

[11] Uenishi,Y., Masui, F. and et al.: Trend Information Extraction Based on Relative Expression Participated on MuST T2N Subtask, in [8], pp. 509-514, 2008.