人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-01

表形式でインタラクティブにクラスタリングを行い，可視化するツールの開発と実践

伊藤貴一[1]　白土由佳[2]　熊坂賢次[3]

Takaichi Ito[1], Yuka Shiratsuchi[2],Kenji Kumasaka[3]

[1]慶應義塾大学院政策メディア研究科
[2]産業能率大学
[3] 慶應義塾大学環境情報学部

Abstract: 多数のアイテムとその関係を可視化するために，縦の列は同一クラスタ，横の行は同じくらいの頻度という表形式で Web ブラウザ上に表示するツールを開発した．このツールは，クラスタリングもでき，UI の操作により分析者の考えを反映する制約付きクラスタリングを可能にしている．このようなツールを開発し，実際にデータを分析した．

1.はじめに

インターネットが社会に浸透するに従い、多くの人が Blog や SNS を使うようになり、人々のライフスタイルを Web サービス上に表明するようになってきている。このような状況下、社会調査も、アンケート調査をするというものだけではなく、Web にある人々の声を拾う、ソーシャルリスニング[1]が一分野になってきている。ソーシャルリスニングのためのツールが求められている。

このようなツールは、あらかじめ明確な答えがないため、探索的アプローチになってしまう。そのため、探索を支援するツールであるべきだ。この探索のためには、機械処理の結果を見せるだけのシステムではいけない。人間の背景知識や分析意図を結果に反映させるようなものでなくてはならない。そのため、インタラクティブ性は重要であり、人間とデータと機械処理が融合するような、知的インタラクティブシステム[2]である必要がある。

2.ツールのコンセプト

この論文のツールで扱うのは、バスケット分析の可視化である。商品の購買履歴や、自然言語を形態素解析後のデータを用いた分析である。共起関係に基づきアイテム間の関係を可視化する。これにより、商品購買なら、購買の関係図、自然言語なら、言葉の関係図を作り、データにある構造を読み解くことができるというものである。

このようなデータを分析するために筆者は、縦列をクラスタ、横列を頻度のレイヤーの二軸を使い、表形式でインタラクティブに表すツールを作成した。これは、第二著者の論文において社会分析に使われた手法[3]のツール化である。

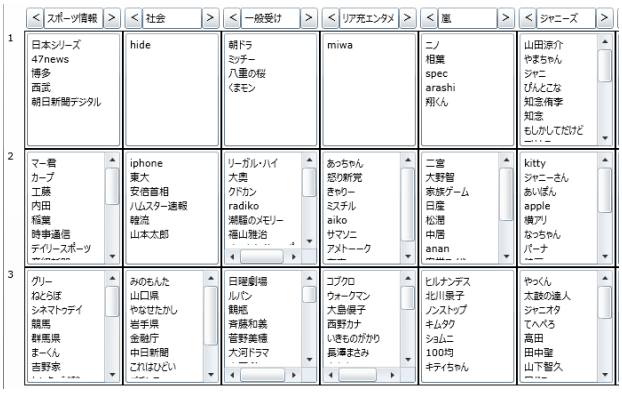


Fig.1

最終的には、Fig.1 のような可視化を行う。
以下、作成したコンセプトを示す。

2.1 べき乗分布と頻度の層（レイヤー）

商品購買履歴のようなバスケット形式のデータを分析するとき必ず発生するのは、少数の高頻度のものと、多数の低頻度のもので構成されるべき乗分布になることである。自然言語処理の世界では、ジップの法則[4]と呼ばれるものであり、マーケティングの世界では、ロングテール[5]と呼ばれるものである。このような分布は必ず発生するものとして、分析に予め組み込む必要がある。べき乗分布の性質として、両対数グラフを作ると線形に近似するというものがある。（べき乗分布は反比例に近似し、ｘｙ＝α を両対数にすると、log(y)=−log(x)+log(α）となり直線となる。）これを利用して、最大の頻度の対数と、分析に使う最小の頻度の対数の差をとり、それを等分割することで、頻度の層（レイヤー）を作るとい

うことをする。これは、上のレイヤーからピラミッド状にアイテムの個数が増えていくものとなる。これを行にして、上の行は頻度の高いものであり、下にいくと頻度が小さいものと、直感的にわかるものとした。

2.2 インタラクティブな関係の表示とクラスタリング

関係性の表示のために、アイテムのクリック時、関係が強いアイテムに色を付けるということをしている。これは、グラフにおける、エッジを、インタラクティブに見せていることに相当する。そのため、複雑な模様になってしまいがちなネットワーク図と同じ情報をすっきりと見せるようにしている。縦の列で、なるべく関係の強いもので固めるという形で、クラスタリングを行う。教師なしで、関係のみを用いで行うため、これは自己組織化させているともいえる。

Fig.2

Fig.2 において、赤は選択したアイテムであり、橙色は、関係しているアイテム。薄い橙は、弱い関係である。ネットワークを可視化しているといえる。

2.3 概念化とメタ認知

このようなクラスタリングによる、データの可視化だけでは不十分である。社会学的分析にするためには、概念化が必要である[6]。概念化とは、分析の全体像を考えることであり、それにしたがって、クラスタを表す言葉を探し、名付け、その塊を分析者が把握することである。そのような概念化をすることで、事象の解釈が可能になる。この行為は、認知科学的に言えば、メタ認知による言葉での外化である[7]。そのため、クラスタに名前をつけられるようにしている。

2.4 制約付きクラスタリング

名前をつける時に困るのは、データに忠実で機械的な処理に基づくクラスタリングの結果では、人間が考える概念とは、しばしばズレることである。このズレを解決するためには、人間の背景知識をクラスタリングの結果に反映する制約付きクラスタリング[8]という手法を用いる。制約付きクラスタリングとは、MustLink、CanNotLink を予め指定し、その情報を付加した上でクラスタリングする手法である。ここでは、高間[9]のように、グループ指定による、制約情報を UI の上で加えるようにする。ユーザが直感的に行えるように「固定する」というメタファで説明している。UI により固定化したアイテムは、クラスタリング時には、動かなくなる。当然、動かないとしても、そのアイテムの関係情報は使って、他のアイテムには影響している。

2.5 クラスタリングの失敗の可視化

このクラスタリングは、必ずどこかのクラスタに所属させるハードクラスタリングのため、うまくクラスタスタリングできていないアイテムが発生することがある。このようなものは、複数のクラスタと関係をもつものであり、ネットワーク構造的にはハブである。ネットワーク分析ではハブの重要性はしばしば指摘される。特に、低頻度のハブは、KeyGraph で言う、赤ノードに相当し、KeyGraph の考案者である大澤の主張では、そのようなものにはチャンスが眠っているとされる[10]。このようなクラスタリングに失敗しているハブ的なものは、縦列に並んでいるものは、同じクラスタであるという、可視化のルールから外れるため、それには、赤丸をつけ可視化する。

2.6 クラスタ間の関係の可視化

縦と横の二次元の可視化では、クラスタ間がどのような関係になっているかがわからない。概念化により、概念同士の関係がどのようになっているかを知るためにも、クラスタをノードとして、グラフとして可視化した。この際、クラスタに対して適切な名前を与えていないと、機械的な名前になり、イメージ出来ないものになってしまう。このことでも、名前をつけることを促進させている。

2.7 属性情報の付加と可視化

データには、いつどこでだれがといった、５Ｗ１Ｈの情報が本来的にある。このような情報をテーブルに重ねあわせる仕組みを用意する。

3.実装

実装は、C#で行い、Silverlight というブラウザのプラグイン上で実行できるようにした。そのため、Windows と Mac のブラウザ上で実行できる。

Silverlight にしたのは、Windows と Mac 両方で実行できるということと、最新版への更新が簡易なこと、ブラウザ実行とはいえ、ローカルファイルを扱え、通常のアプリケーション同様のことができるからである。次のアドレスで公開している。
http://goo.gl/VuHWa6

3.1 入力ファイル

リレーショナル・データベースからの出力を入力に仮定している。そのため、入力ファイルが１つでは済まない。すべては、UTF8 でエンコードされたテキストファイルでなければならない。

- 構造用ファイル
- UserId と ItemId のデータ
- UserId と属性のデータ

これらの TSV（タブ区切りデータ）が必要である。さすがに３つのファイルを用意するのは大変なので、一つで済むような仕組みを検討している。

3.2 画面の説明

ツールの基本画面はこのようになっている。(Fig.3)

Fig.3

画像に振った番号にそって機能を説明すると、

1. ファイル関係。ファイルの入出力を行う。
2. 表示設定。表の大きさなどの設定。
3. アイテムのクリック時、表示する関係の数と指標の設定。
4. マウスのモード設定。デフォルトでは、選択であるが、移動に変更すると、アイテムが移動可能になる。削除にすると、削除できる。
5. 固定化モードの ON/OFF。ON にした時、アイテムの横にチェックボックスが現れ、チェックしたものは、クラスタリング時に動かない。制約付きクラスタリングための制約を与えることができる。
6. クラスタリングパネル。クラスタリングの設定と実行、経過の表示を行う。

結果画面は Fig.4 のようになっている。

行が、頻度のレイヤーを示し、上から頻度が大きいものから並んでいる。列が、クラスタを示し、基本的に、塊を形成している。列の上には、自分でクラスタの名前が書き込めるようになっている。また、列は左右に移動できるようになっており、解釈に最適な並びを探索することをできるようになっている。

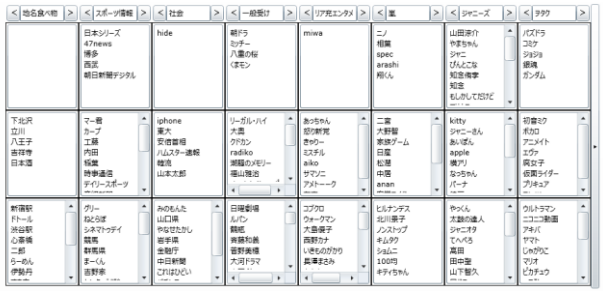


Fig.4

3.3 クラスタリングのアルゴリズム

クラスタリングのアルゴリズムは、可視化に合わせて作成した。K-Means 法の改変である。
クラスタ数は予め UI で指定する。固定化アイテムも予め指定してもいい。

1. すべてのアイテムを頻度レイヤーに沿ってランダム配置する。固定化されたアイテムは別である。
2. すべての非固定化アイテムにおいて、指定個数分の補正信頼度の高い順にアイテムを抽出し、それぞれのクラスタごとに補正信頼度の平均をとり、もっとも高いのを勝者クラスタとし、そこに移動させる。（ただし、移動は同一頻度レイヤー間で行い、移動情報は一時表に保存）
   (ア) 補正信頼度は、信頼度-支持度。Lift 値が割り算である代わりに引き算である。Lift 値は、割り算を使うので、値域が０から無限大までとるが、補正信頼度は、-1 から１の間で実データでは、-0.1~0.3 ぐらいの値に収まるため、扱いやすい。
3. 一時表に保存したものを本表にアップデート。固定化アイテムは同じ位置のままである。
4. 移動したアイテムが０なら終了、あるなら、2 に移動。
5. 1~4 を、指定回数繰り返し、評価値が最も高いものを表示させる。

このようなアルゴリズムにした。重み付きグラフのクラスタリングである。

3.4 クラスタリングの評価指標

クラスタリングの評価指標にはジニ係数を使う。ジニ係数は、格差を示す経済指標として有名だが、機械学習でも使われている。（例えば決定木[11]）。性

質としては、値の格差が大きいと１に近づき、格差が小さいと０に近づく。このジニ係数を使い、２つの軸を持って評価する。

1. すべてのアイテムが縦列で、まとまっていることがいいクラスタリングである。
   (ア) すべてのアイテムで、指定個数分の補正信頼度の高い順にとり、それをクラスタごとに総和を求める。これを変数としてジニ係数を求める。
   (イ) 求まったそれぞれのジニ係数の平均値を出す。１に近いほどいい。
2. 可視化として、それぞれのクラスタに入っているアイテムの数が均等に近いのがいいクラスタリングである。
   (ア) 頻度レイヤーごとに、クラスタごとのアイテム数を数え、それを変数としてジニ係数を求める。
   (イ) 求まった各頻度レイヤーごとのジニ係数の相乗をだす。０に近いほどいい。
3. １と２の２つを掛けあわせたのを最終的な指標とした。ただし、１と２は向きが違うので、向きを揃えた。

１は、ツールでアイテムそれぞれをクリックした時、関係しているアイテムが表示されるところが、縦の列でなるべくまとまっているというのを表現している。しかし、１だけでは、不十分だった。この指標を最大化するには、クラスタの空欄を増やせば増やすほど高くなるため、クラスタ数の指定が意味を成さないことが判明した。そのため、なるべく空欄をださないように、それぞれのクラスタに均等になったものがいいクラスタリングと評価されるように２の指標を追加した。

3.5 クラスタリングの失敗の検出

１の評価指標は、一つ一つのアイテムにおいて、１に近づけば近づくほど、クラスタとしてまとまっていることを意味する。逆に、０に近いものは、複数のクラスタと関係を持っているアイテムであるということ、すなわち、ネットワーク構造上、ハブになっているものだと思われる。縦の列でクラスタを作っていることが可視化のルールなので、そのルールから外れているので、このようなものを可視化する。具体的には、１の平均値を求める前のデータで、ジニ係数が低い物順に指定個に対して、赤い丸をつける。

3.6 クラスタマップの作成

表形式では、クラスタ間の関係がわからない。そのため、クラスタをノードとした、グラフを作成した。クラスタ間の関係には Jaccard 係数を使い、エッジの足切りには、Lift 値を使った。Lift 値の足切りにより、エッジの数を増やしても、グラフは完全グラフにはならないようにした。また、ノードの名前は、名付けていないと、素っ気ない機械的な名前にすることで、積極的にクラスタに名前をつけることを促進している。 また、関係性の実データが見えるようにもしている。

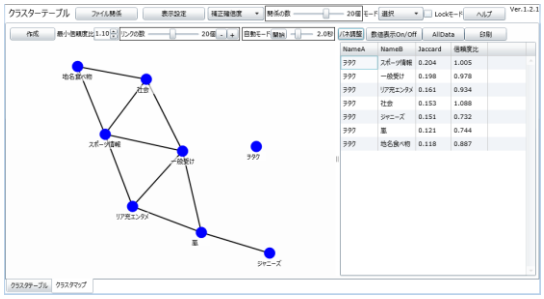


Fig.5

4.分析事例

2013 年、テレビドラマの「半沢直樹」は、最終話の視聴率が 42.2%（ビデオリサーチ調べ）という空前のヒットを飛ばした。この「半沢直樹」についての調査を行った。

分析データは、ツイッターで、「半沢直樹」の公式アカウント(@Hanzawa_Naoki)をフォローしているユーザ（45,315 人）のツイートを 2013 年 11 月に取得した。クリーニングとして、オープンであり、言語が日本語であり、ツイート数が 2000 以上のユーザを使った。約 11000 人に絞られた。そのツイートの中から、形態素解析を行い、頻出語 250 語を抽出し、その頻出語を用いて、バスケットを作成した。

また、そのユーザが特につぶやいた「半沢直樹」の俳優名と役名を属性とした。

このデータを用いて、本ツールを使い分析を行った。

まず、はじめにクラスタリングを行った時、ジャニーズと嵐が混ざった感じの大きいクラスタを形成していた。この２つは当然結びつきが強いが、数として大きすぎるので、２つを分けるために、次のように固定化行った。嵐のメンバーと、中居くんなどの他のジャニーズのメンバーを分けるようにした。

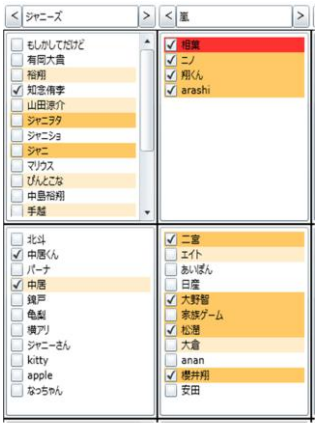


Fig.6

Fig.7

最終的には、クラスタに名前をつけて、Fig.7 のような結果になった。「半沢直樹」はテレビドラマであり、テレビドラマ的な要素が大きいことがわかる。その中でも、「おっさん向けテレビ」の大河ドラマ、朝ドラが好きな層と「若者向けテレビ」の若手のお笑いタレントが集まるクラスタと、「嵐」「ジャニーズ」のクラスタが発生したことがわかる。

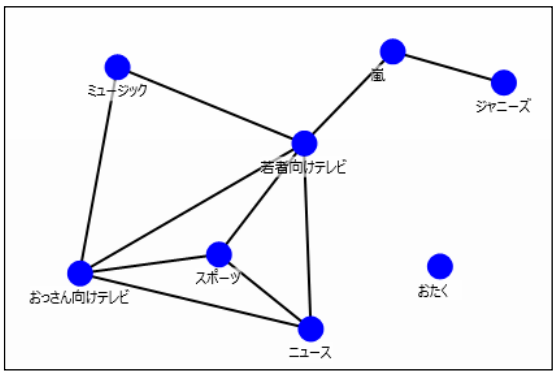


Fig.8

クラスタマップを作るとこのようになる。全体がつながるようにと Lift 値 1.1 で作成した。しかし、「おたく」のクラスタは、テレビ関係のクラスタとは強い関係性はなく、独立関係となった。ジャニーズも嵐を媒介項として全体像とつながっていることがわかるのも面白い。

次に、属性で見る。これらのクラスタは、どの俳優役名と関係が強いか？をみる。赤いところが、その俳優で特化しているところで、青色が特化していないところである。堺雅人、壇蜜、大和田常務の結果を示す。主役の半沢直樹は、タイトル名であるため、分析にはそぐわない。

Fig.9 堺雅人

Fig.10 壇蜜

Fig.11 大和田常務

主役を演じた、堺雅人が特化しているのは、ドラマクラスタだとわかる。一方、女優の壇蜜が、特化しているのは、ニュースクラスタであり、普段ニュースについてつぶやいている層に壇蜜は受けたというのが想像できる。当然、ジャーズクラスタには不人気である。

クラスタマップにおいて、独立だった、「おたく」クラスタのみ特化していたのは、大和田常務である。大和田常務は、その顔芸がネットでヒットして、ネット上のまとめサイトで、いろいろな形でまとめられており、その影響だと思われる。

このように考えると、ドラマ「半沢直樹」は、国民的ヒットなっていったことは、おぼろげながら見えてくる。つまり、普段からドラマを見ている人たちを惹きつけ、ジャニーズ出演でジャニーズ好きな人たちを惹きつけ、壇蜜で、普段、ニュースをつぶやいている人たちを惹きつけ、大和田常務で、テレビドラマを見ない、ネットだけを見ている層を惹きつけることに成功したことが、大ヒットに繋がった、ということが推察される。

このような分析ができるのも、このツールだからこそのものである。

5.議論

クラスタリングは、ランダムな初期配置から作成していくアルゴリズムであるため、同じデータであれば、同じ結果を必ず保証するものではない。また、クラスタリングとしては、大雑把なクラスタリングのため、精密なクラスタリングのために、制約付きクラスタリングの枠組みを利用しているものとなっている。クラスタリングには正解はないとはいえ、分析者の能力に依存するところが大きい。

知的インタラクティブシステムのために、システムとして最小のユーザフィードバックで済むことが望ましいとされる[2]。このツールの場合、機械的な仕組みとして、最小のユーザフィードバックをサポートするような仕組みは存在しない。しかし、意味的な側面と可視化としてのサポートはある。それは、クラスタに名前をつけるのだから、意味合いとして大きい高頻度のアイテムを固定化すべきという意味的な要請と、アイテムをクリック時の関係の表示で、すでに相互に関係があって塊を形成しているものに対して固定化をしてもナンセンスであるということである。そのため、固定化すべきものは、UI 的におぼろげながら示していると言える。しかし、これも、分析者の能力への依存が大きく、初めて使うユーザにとっては不親切であり、何かしらの改善の余地はあるだろう。

とはいえ、ツールの使用者に聞くと、初めに出力される結果にある程度満足してしまうようである。そのため、固定化による制約付きクラスタリングは、アドバンストな機能であるといえる。しかし、このような手段が存在するかしないかでは、分析者の選択肢が増えることなので有効である。

本論文で示した、インタラクティブなツールは使わないと良さがわからない。ツールは Web ブラウザ上で動かせるので、ぜひ使って分析してみて欲しい。

参考文献

[1] 萩原 雅之,次世代マーケティングリサーチ, ソフトバンククリエイティブ, 2011

[2] 岡部正幸 山田誠二,知的インタラクティブシステムにおけるインタラクションデザインとは何か,2013,JSAI

[3] 山崎 由佳,熊坂 賢次,共有化と生活化から生成される2つの"かわいい":4ファッションスタイルをめぐるネットコミュニティ分析 ファッションビジネス学会論文誌 1348-9909 ファッションビジネス学会 ,2012

[4] Zipf, G. K, The Psycho-Biology of Language, Boston-Cambridge Mass. Houghton Mifflin.,1935

[5] Chris Anderson, "The Long tail," Wired, 2004.

[6] 熊坂賢次, 山崎由佳,"ソーシャルな時代・柔らかい構造化手法 そしてライフスタイル論", AD・STUDIES, Vol.40 Spring, 2012.

[7] 諏訪正樹 身体知獲得のツールとしてのメタ認知的言語化, 人工知能学会誌, Vol. 20, No. 5, pp. 525-532..2005

[8] Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge." ICML. Vol. 1. 2001.

[9] 高間康史, 三宅遼祐,グルーピング操作に基づくインタラクティブなクラスタリング対制約生成手法の考察, 第 27 回人工知能学会全国大会, F4-OS-04-3,2013

[10] 大澤幸生 チャンス発見の情報技術、東京電機大学出版局,2003

[11] L.Breiman, J.H.Friedman, R.A.Olshen and C.J.Stone, "Classification and Regression Trees", Wadsworth ,1984

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-02

文書情報を活用した連想支援システムの開発

Development of an “association” support system using document data

荒井　豊文 [1]

Toyofumi ARAI[1]

[1] 中京大学
[1]Chukyo University

概要： 蓄積した文書情報の中から，ユーザーが指定した情報要求に基づき抽出した情報を木構造(語木)表現で視覚化し，さらにユーザーの操作に応じてインタラクティブに変化させることで連想を支援するシステムを試作した．

文書情報の木構造表現による視覚化や，視覚化した文書情報を変化させる動作には，人のメンタルモデルに関する先人らの研究により得られた知見や経験則を反映させることを試みた．

人の情報認知に関するメンタルモデルに則した動作で情報の提示を行えるようにしたことで，連想支援に有効な効果が期待される．

Abstract: We have created a system that provides support for association by visualizing in a tree structure (word tree) information filtered based on a request for information specified by the user from information stored in documents and, further, based on user operation, changes this information interactively.

Through the visualization of document information in the form of a tree structure and the operation of changing the visualized document information, we have tried to reflect the knowledge and rules learned through experience obtained through the research of our predecessors into the human mental model.

It is expected that, by presenting information through actions that follow a mental model of human information recognition, that there will be useful benefits for association support.

１．はじめに

研究など創造的要素を含む知的活動においては，新たな気づきや発想を得るためのアプローチとして，連想を用いることがある．連想の情報源として論文などの文書情報を用い，これを熟読することが多い．しかしながら，文書テキストのままでは内容理解のための認知的負荷が高く，文書情報から連想を行う上での障壁となっていることが予想される．

そこで，文書情報理解のための認知的負荷を低減させ，連想を促し，気づきや発想を生み出しやすくすることを狙った支援システムを検討，試作した．

連想支援の方策としては，非定型情報である文書情報に対し一定のルールを適用し，形をもたせ視覚化し，視覚化した情報をユーザーの操作に応じて変化させることができるようにすることで，ユーザーの情報認知に刺激を与え連想を促す情報提示システムを考えた．

このようなシステムにおいて文書情報に形を持たせるためのルールは，人のメンタルモデルに則した方法を適用するのが認知的負荷の低減に有効と考える．またシステムとユーザー間のインタラクティブな情報のやり取りにおいては，行った操作に応じてシステムが提示する情報の変化を視覚的に認知できることも有効と考える．そこでこれら２点の実装に重点を置きシステムを試作した．

２．システムの検討

検討したシステムの構造を図１に示す．連想のもととなる文書情報を格納した情報源と，その情報をハンドリングするロジックからなる構造とした．ハンドリングロジックではユーザーの要求に対応した情報を情報源から抽出し，予めユーザーが指定した描画方法で視覚化し提示するとともに，一旦提示した情報に対し，ユーザーが操作を加えることにより

視覚情報を変化させることができるようにした.

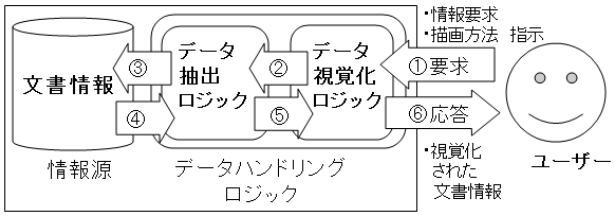


図１．システムの構造

2.1 情報源

情報源に用いる文書情報には特許情報を用い，これを関係型データベースに格納して用いた．特許情報は先人たちの知的活動の成果物であり，しかも電子化された情報を大量かつ容易に入手し利用できることから，提案システムの有効性を検討する際にも適切であると考えた．またデータベースを用いたのは，文書情報を解析に適した形に構造化してストックしておくことで，これを利用する様々な要求に対応できるようにするためである[1].

2.2 データハンドリングロジック

データハンドリングロジックは，データ抽出ロジックとデータ視覚化ロジックに分離し開発した.

2.2.1 データ抽出ロジック

データ抽出ロジックの中心は情報源である関係型データベースへのアクセス機能であり，データ視覚化ロジックでユーザーが指定した情報に基づき SQL 文を生成，実行し，必要な情報をデータベースから抽出する．さらに，抽出した情報をユーザーの指定した条件に応じて加工，編集する.

2.2.2 データ視覚化ロジック

システムとユーザーとのインタフェースであり，本システムの最重要部分である．メンタルモデルに則して情報を視覚的に表現するにあたっては，シンプルなルールにより意味づけした形で提示することとした．視覚情報とその意味との関係が複雑になると，新たな認知的負荷がユーザーに生じる恐れがあると考えたからである

また，ユーザーの思考を中断させることで新たな認知的負荷が発生させることの無いように，一連の操作が容易に繰返し実行できることも重要と考えた.

３．システム開発内容

3.1 機能実現の方策

情報の視覚化や，システムとユーザーのインタラクティブなやり取り，大量文書情報中からの情報抽出などに有効と思われる，人のメンタルモデルに関する先人の研究成果や経験則には，たとえば,

① 人は情報の集まりを見ると，そこに含まれる規則を見出そうとする．[2]
② 人は情報の並びを見ると，そこに含まれる規則を考え，それを元に次に現れる情報を先読み(予測)しようとする．[3]
③ 段落など，特定の部分を単位として検索し提供することにより，関係した情報を効率的に抜き出すレレバントな情報検索ができる．[4]
④ 複雑な理論により少量の情報を分析するよりも，単純な理論で大量の情報を分析した場合の方が有効な結果が得られる場合がある.

などがある.

①，②よれば，情報の提示方法を工夫することで連想が促進できると考えられる．また，③は，大量の情報の中から有効な情報を抽出すことに関するするものであり，人は文書中の纏まった箇所に特定の話題に関する内容を集中させる傾向があることを示すもので，これを利用すれば有効な情報をユーザーに提供し易くなることが考えられる.

さらに，これらのほかにも，ユーザーが使って楽しく感じるか否かということもシステムの有効性に影響することが知られている.

そこで，先人らの研究成果や知見や経験則を参考とし，検討した結果,「連想ゲーム」[5] 的動作を実装することとした．連想ゲームでは回答者が正解を答えるまで，ヒントとなる言葉が繰り返し提示され, 正解に至ることが必要であるのに対し，提案システムでは正解は求めない．またヒントに相当するものとして提示される情報は，抽出された文書情報に含まれているものに限定される．こうした違いはあるものの，関係を持つ情報を次々と提示し連想に結びつけるといった基本動作においては共通するものがあると考えた.

3.2 ユーザーへの情報提示

抽出した文書情報を視覚情報としてユーザーに提示し，さらに「連想ゲーム」的動作をさせるための単純化したイメージを図２に示す.

描画の形とその意味の関係の基本ルールとしては，ユーザーの情報要求に合致しているとして抽出された文書中に含まれる語の，出現頻度の多さを円の大きさで，またその文書中に含まれる文中での各語の共起関係の強さを円どうしを結ぶ線の長さで示すとした．これにより，根語を始点とし

て線で結ばれた各語をたどり終端となる語までの一連の語の並びが一つの文に相当し，円で示された語は文中に出現する語群を表すようにした．

根語に用いる語は，ユーザーの情報要求として入力された文中に含まれる語，もしくは情報要求に合致しているとして情報源から抽出された文書内の特徴語である．いずれを根語に用いるかは，ユーザーが指定できるようにした．

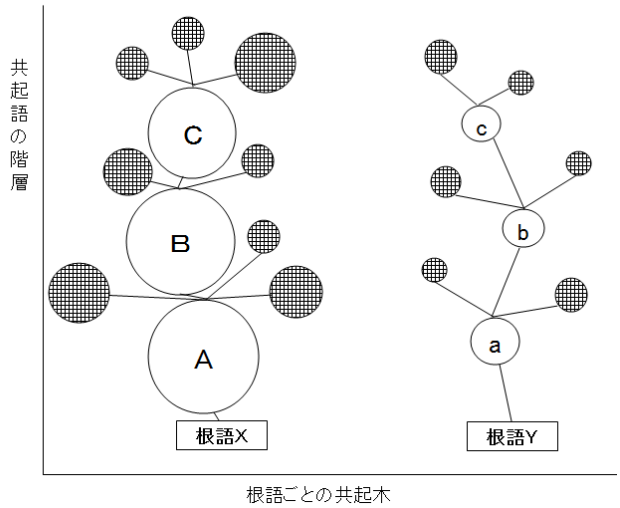


図２．語木表示イメージの例

情報要求に含まれる語を根語に用いた場合，語木はユーザーの視点を反映させたものとなる．一方，特徴語を根語に用いた場合はユーザーの視点とは別に，情報要求に合致した文書が持つ特徴をもとに形成した語木となる．前者のように視点をもって情報を見ることも重要ではあるが，後者の機能により，より自由な連想の元となる情報の提示が期待される．

このような表現ルールを用いた「連想ゲーム」的情報提示の主な動作は，

・ユーザーの操作に応じ，根語から枝葉語が展開するように段階的に表示する．
・語木を構成する個々の語系列単位で順番に強調表示したり，ユーザーが任意の一語を指定することで，その語が含まれる系列を強調表示する．
・語木で表現した語系列情報を元に再検索するなど，次の操作を連携して行うことができる．

などができるようにすることとした．

情報の関係を視覚的に表現することに関する先行研究では，ネットワーク図で表示するものが多い．提案システムで情報の視覚化に語木（木構造）を用いたのは，「大きさ」「長さ」「始点と終点」「方向性」「並び」「順序」等，人が容易に認知できる情報の尺度を対象に持たせることで，メンタルモデルを利用する効果をより有効にし，他の視覚化方法よりも情報認知において優位とすることを狙ったからである．さらに語木で表現することにより，人が「木」に対して持っているメタファーが連想の促進に生かされる効果も期待した．

このような表現方法によれば，たとえば文書中で強調されている内容については，同様の語群の語を用いて繰り返し文書中に記述されるであろうことから，それら語の出現頻度，共起頻度ともに高くなると予想され，図２の左に示した根語Ｘから始まる「X-A-B-C」の語系列のように語を囲む円が大きく表され，また語どうしを結ぶ線が短くなり互いの語が近くに描画されると推察される．逆に，述べられる頻度が少ない文を構成する語群は，図２で右に示した根語Ｙから始まる「Y-a-b-c」の語系列のように語を囲む円が小さく表され，また語どうしが離れて描画されると推察される．

さらに，複数の文書や段落の情報を一つの文書や段落とみなして描画することもできる機能も付加した．これにより，たとえば作成者の異なる複数の文書に含まれる情報を用いて描画した語木において，大きな円で囲まれた語の並びの語系列が出現した場合には，複数の人により同じ主張がなされている可能性があると推察され，より信頼性の高い情報を示すものになることが予想される．

マウス操作で根語から順に枝葉となる語を表現する「連想ゲーム」的動作では，各操作を実施する時点までに描画されていた語もしくは語群が，次に描画される語を推測するヒントの役目を果たす．動作イメージを図３に示す．

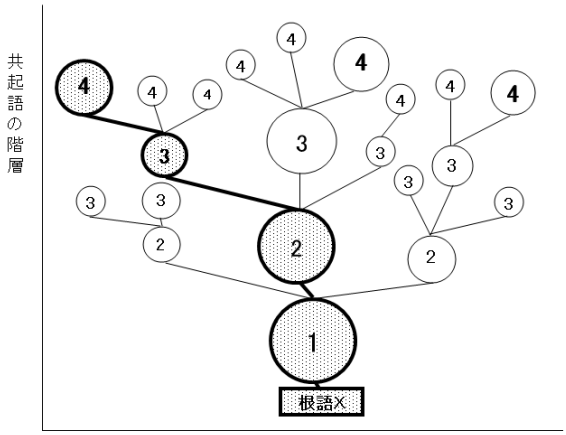


図３．段階的な語木表示及び特定語系列強調の例

図３の円の中に記した数字は段階的に描画される順番を表し，たとえば①の語をクリックすると②で表された語が表示される．

さらに，ユーザーの操作により，図３に太線で示したように特定の語系列を構成する語を他の語系列

と区別し強調表示することもできるようにした．さらに，強調表示した語系列に含まれる語を用いて文書データベースを再検索し，語木を再描画することや，該当する文書の原文を検索できる機能とも連携させた．これにより，ユーザーの思考を中断させずに必要な情報を提示できる効果も期待される．

なお，情報要求と，情報源から抽出する文書や段落との適合性の評価には一般的な Tf･idf 値を用いた．描画する円の直径の計算には前述したように相対出現頻度を，語間の関係の強さを示す語間の描画距離は共起頻度の相対値の逆数を用いて計算した．

情報要求に基づき文書データベースから情報を抽出する際に指定できる抽出対象情報の単位と，語木描画時に指定できる条件項目を表１に示す．

表１．情報抽出および描画に指定できる条件

■データベースからの描画情報抽出単位
・文書単位 ・段落単位
■抽出した情報の語木描画時に指定できる条件
・描画対象情報の単位 （特定文書／特定段落／複数文書／複数段落） ・共起頻度下限値 ・語木描画階層数 ・共起分析対象 （文内共起／段落内共起／文書内共起）

４．システム動作確認テストと考察

4.1 準備

4.1.1 テストに用いた文書情報

動作テストに用いた文書情報は，前記したとおり特定技術分野の公開特許公報(以下，「特許広報」)を特許庁特許電子図書館よりダウンロードして用いた．

用いた公開特許公報の数を表２に示す．

表２．テストに用いた特許情報

入手先	特許電子図書館（ＩＰＤＬ）
入手日	２０１０年５月３１日
入手件数	４３２件（公開特許公報）

4.1.2 テスト用文書データベースの作成

特許公報中のテキストを形態素解析し，名詞，動詞，形容詞のみについて出現形，基本形及びその語が出現する文書,段落,文等に関する情報を格納し，文書データベースとし，情報源に用いた．

文書データベースに格納した文の数，段落数，語数を表３に示す．432 件の公開特許公報から約 200 万語が抽出でき，これら全てを格納した．

表 3. テストに用いた特許情報の段落数，文数，語数

段落数	109,258
文数	139,570
語数	1,996,342

4.2 テスト結果と考察

4.2.1 基本動作

動作確認テストは情報要求を「地球温暖化防止のための二酸化炭素ガスの分離除去」とし，また，語木の根語は，情報要求に基づき抽出された文書中の特徴語とし，共起度下限値等描画条件を変え，意図したとおり語木が表されるか，また，語木を形成する語間の関係が描画できるかの動作を確認した．

抽出された特定の文書について，共起度下限値を 11 とし描画した結果を図４に示す．語を囲む円の大きさで相対出現頻度を，語間の距離で共起頻度が表現できてはいるものの，条件を変えて繰り返し実施した結果，共起度下限値を小さくした場合，描画される語が増えることにより語木が「混んで」しまい，情報が読み取れない状態となった．

そこで，描画した語系列を一覧リストとして表示する機能を追加した．これにより，語木中で任意の語系列を選択すると，それに対応しリスト中の文字列も強調表示される．また描画した語木中で任意の語を指定すると，その語が含まれる全ての語系列を強調表示する機能や，語木を段階的に表示する機能等，操作に応じて描画内容を変化させる一連の機能が意図した通り表示ができることを確認した．

4.2.2「連想ゲーム」的(段階的)表示機能

「連想ゲーム的」的機能の実装例として，語をクリックすることで共起関係にある語を段階的に表示させる機能を，図５に根語「フロン」についての動作例で示す．

根語「フロン」(①)から順番に，「分解」(②)，「光」(③)，「反応」(④)を選択してゆく経過を示す．順番に語をクリックするたびに共起語が次の階層の候補語として赤色で表示される．候補語の中の特定の語を選択すると，選択された語以外の語が青色に変わるとともに，次の階層の候補を赤色表示する．この一連の操作でのシステムとユーザーとのインタラクションを通じ，連想促進を期待している．

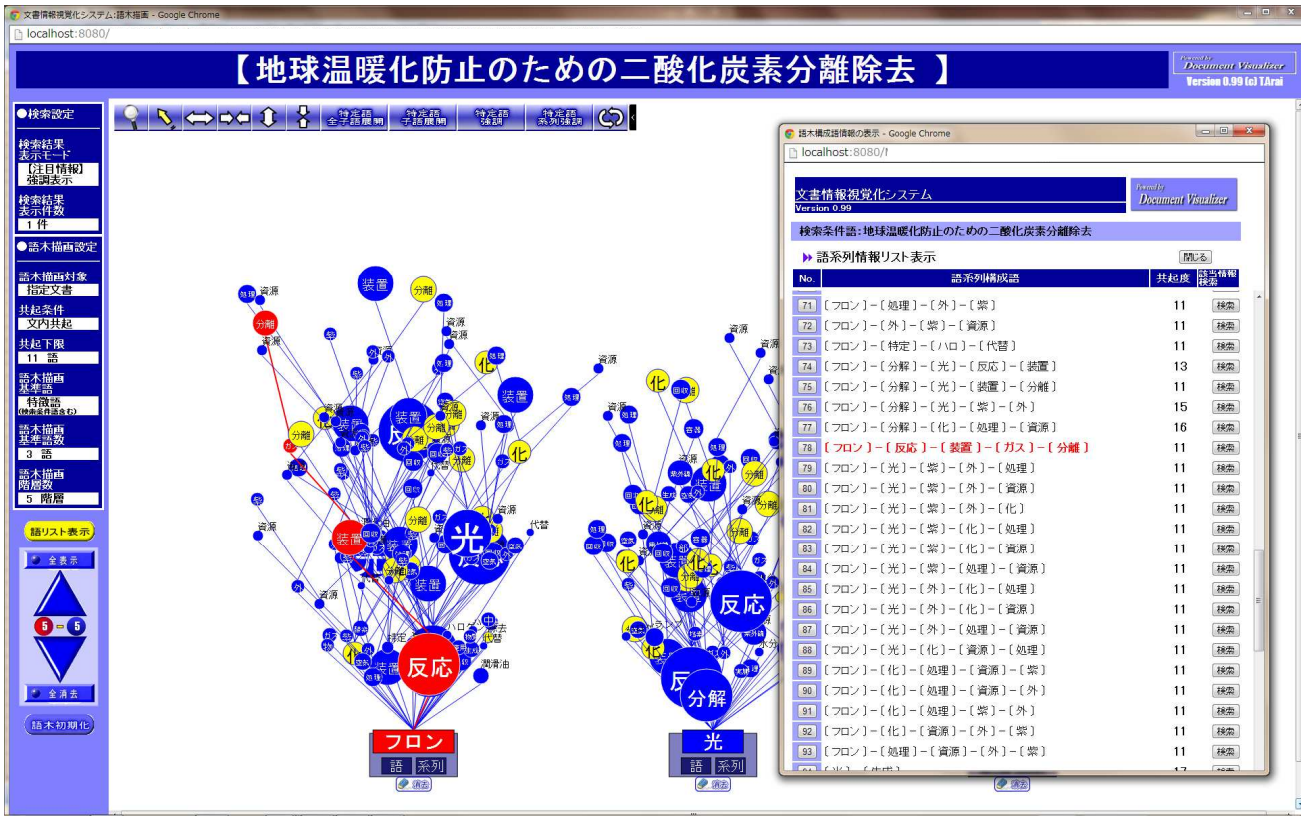


図４．語木の描画と語系列リスト表示例

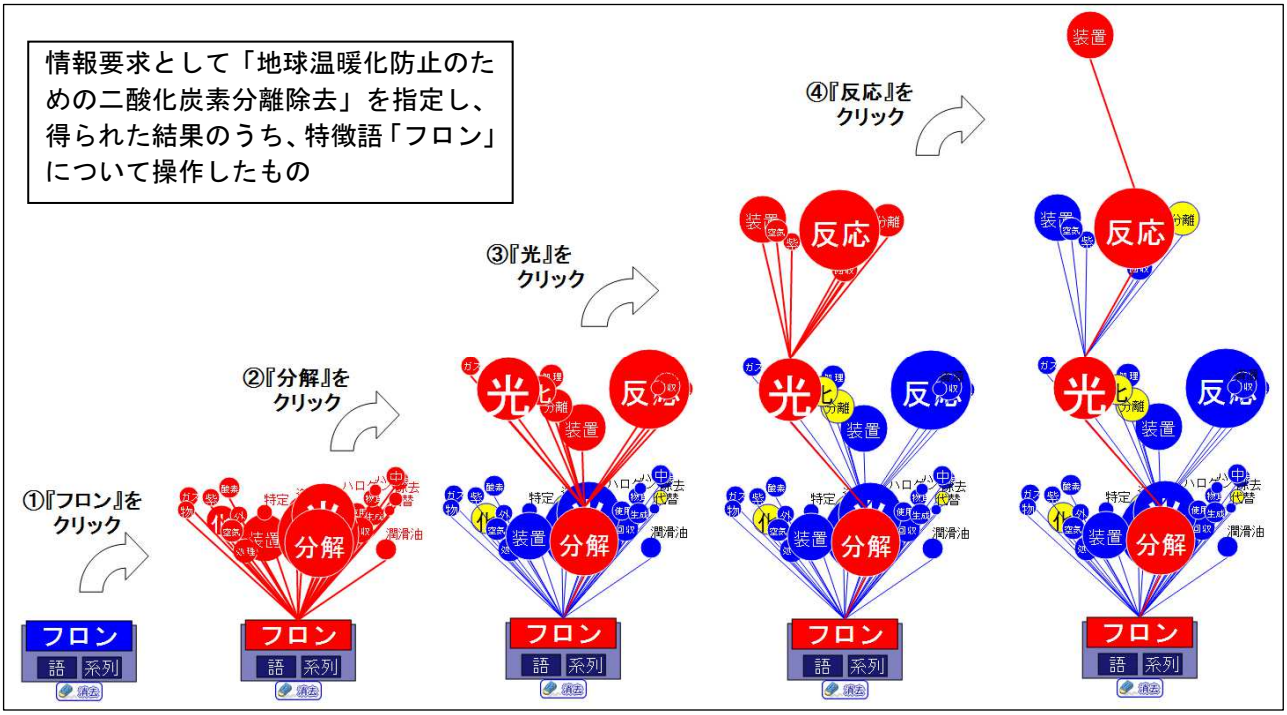


図５．「連想ゲーム」的語系列表示機能の例

4.2.3 語系列の選択と文書検索の連携機能

描画した語木中の指定した語系列に含まれる語群を検索条件として用いて文書データベースを再検索する動作例を図６に示す．図中赤で示した語系列を構成する語が自動的に検索条件語として用いられ(図中「指定語」欄)，該当する語を含む段落や文書が検索できている．この操作では，ユーザーが注目

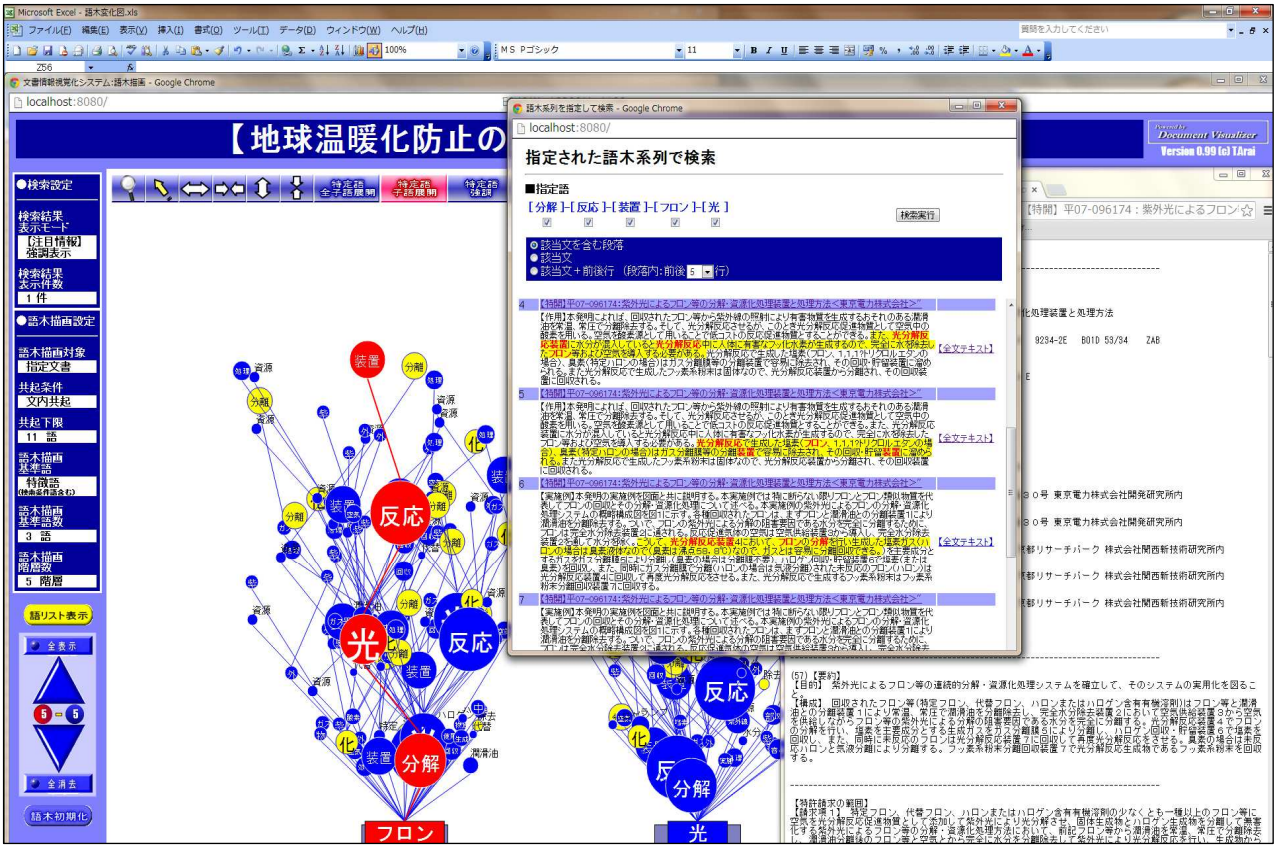


図６．語系列の選択と文書再検索の連携機能の例

した語系列構成語をそのまま検索条件として用いて再度文書データベースが検索できるので，思考の中断を極力抑えることが出来ると考える．

５．今後の展開

まずは提案システムの有効性の検証が必要である．しかしながら検証は，有効性の評価が検証に参加するユーザーの知識や経験に依存するため，検証方法を考案することは非常に困難なことが予想される．そこで，広範な分野の多くのユーザーに開放し，どのような場面でどのような経路の操作をしたときに有効な連想や気づきや発想が得られたかの情報をフィードバックしてもらうことにより，本システムが有効なケースを検証する方法での検証を考えている．

また機能にも改善を必要とする課題がある．まず，本格的な利用に向けては，パラフレーズや語のゆれへの対応が必要である． さらに，ユーザーにより有効な情報を提示するためには，情報源に用いる蓄積文書量を増やすことも必要である．一方，今回の試作および動作テストでは対象文書の分野を絞り，約 400 件程度の公開特許公報を用いたにもかかわらず，文書データベースに格納した語の数は約 200 万語となったことから，蓄積文書量を増やすと処理対象となる語の数が飛躍的に増大することが予想される．たとえば特許庁電子図書館所蔵の全特許情報を提案システムに格納し利用するとした場合，データベースに登録する語数を試算したところ約 550 億語となった．その場合，描画対象情報の抽出，分析時間がユーザーの要望に応えられなくなることが予想される．この問題への対処はデータベース処理を高速化することが必要と考えている．

参考文献

[１] 菰田文男，那須川哲哉，技術戦略としてのテキストマイニング，中央経済社，東京，2014.
[２] D.A.ノーマン，誰のためのデザイン，新曜社，東京，1990.
[３] 酒井邦嘉, 脳を創る読書, 実業之日本社，東京，2010.
[４] 野末道子，上田修一，”論文段落を対象とした日本語全文検索データベースの検索”，情報処理学会論文誌, Vol.1993，No.39，pp.9-16，1993.
[５] NHK(1969-1991) 「連想ゲーム」 NHK アーカイブス NHK Homepage http://cgi2.nhk.or.jp/archives/tv60bin/detail/index.cgi?das_id=D0009010143_00000 (2014，Feb，15)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-03

内部構造解析機能と脚注表示機能を備えた論文閲覧システム

Paper Browsing System with Structure Analysis and Displaying Annotation on Side-note Windows

阿辺川 武 *　相澤 彰子
Takeshi Abekawa　Akiko Aizawa

国立情報学研究所
National Institute of Informatics

Abstract: In this paper, we introduce our on-going efforts to construct a scientific paper browsing system to assist users to read and understand advanced technical content. The paper features on two major functions that are prerequisite for such systems: document structure analysis for image, PDF, and XML formatted articles, and automatic link detection that help users access richer information from diverse external sources. We also present technical details of our current implementation to generate and display the linked external data in side-note windows with a target paper image.

1　はじめに

近年，学術文献の電子化は大きく進み，論文は投稿から出版，読者の手元まで印刷媒体を経由することなく電子フォーマットで流通することが一般的になってきた．一部の海外学術出版社では独自の XML フォーマットを定義し，1 つの XML ファイルから紙の印刷物，あるいは PDF，XHTML，EPUB といった電子フォーマットへ変換する 1 ソースマルチユースと呼ばれる出版工程が実現されている．日本でも科学技術振興機構 J-STAGE 3 において，本文 XML[1] の入稿を推奨するようになるなど着実に XML 化が進んでいる．

論文が XML フォーマットになれば，デバイスの特性に応じて自由にレイアウトや文字サイズを変更できるリフロー型と呼ばれる方法で論文を表示できるようになる．例えば XHTML に変換すれば Web ブラウザで自由な文字サイズで閲覧でき，PMC(PubMed Central) の PubReader[2] のようなデスクトップにもモバイルにも対応した専用アプリケーションで快適に閲覧できるようにもなる．

しかし，多くの学術出版では，電子化といっても紙に印刷可能な PDF ファイルのみを生成し，出版時に PDF ファイルをそのまま配布するに留まっているのが現状である．PDF は画面上でも印刷した紙でも同一のページレイアウトを維持することを目的して策定されたフォーマットのため，紙面レイアウトは常に不変である．そのため画面の小さなデバイスに対して，カラム数を変更したり，文字サイズを変更するなどの表示のカスタマイズが不可能である．電子化が進む以前に発表された論文についても同様に，紙面をスキャンして画像に変換後，PDF にパッケージングするため，やはりレイアウト固定型の PDF フォーマットの論文が数多く流通しているのが実情である．

我々は，Web ブラウザ上で動作する論文閲覧システム SideNoter を開発している．XML 形式の論文のみを対象にできれば，表示レイアウトの自由度の高さから，閲覧に適したインターフェースを開発することができる．しかし，上述の現状では PDF で配布される論文が大半を占めるため，レイアウトの再現性重視のファイル形式をいかに扱うかが課題となる．幸い本文テキスト自体は，電子的に制作された PDF からは比較的容易に抽出でき，スキャンした画像からも OCR を用いて 100%の精度ではないが抽出できる．本システムでは，論文のレイアウト重視の制約を逆に利用し，論文自体は画像で表示し，論文本文から得られる補足的な情報をページレイアウト上に重ねて表示する手法を採用した．

現在，言語処理学会年次大会過去 20 年分の予稿集を用いて，開発中のシステムの使い勝手を検証している [3]．本システムは，PC につながったディスプレイで論文を閲覧するだけでなく，年次大会のような学会の会議に参加し，モバイルデバイスで聴講中の予稿集を閲覧する支援になることもめざしている．

*連絡先：国立情報学研究所
〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: abekawa@nii.ac.jp

[1] https://www.jstage.jst.go.jp/pub/html/AY04S230_ja.html
[2] http://www.ncbi.nlm.nih.gov/pmc/about/pubreader/
[3] http://kmcs.nii.ac.jp/nlp_annual/ で公開中

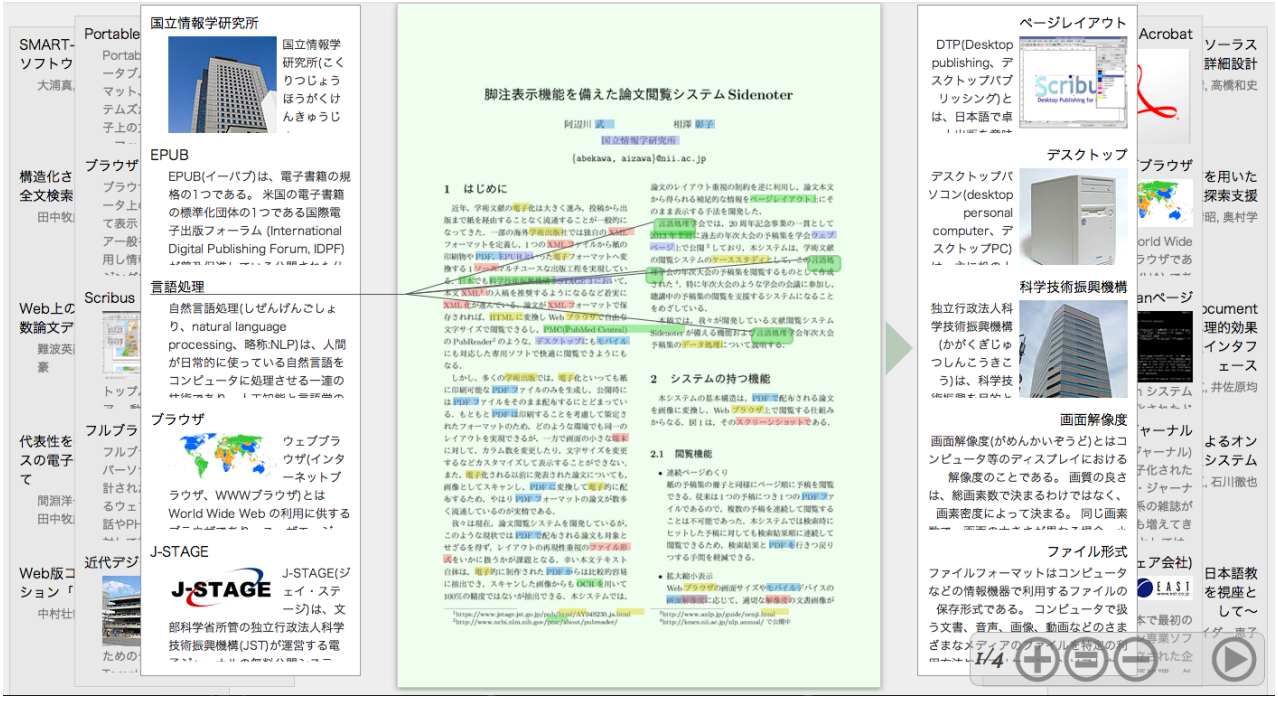


図 1: システムのスクリーンショット

以下，本稿では，我々が開発している文献閲覧システム SideNoter の基本設計およびデータ処理方法を説明し，次にシステムが備える機能について紹介し，最後に外部で付与された注釈を論文レイアウト上に表示する仕組みについて説明する．

2 関連研究

現在，電子的に流通している論文フォーマットの大半は PDF 形式であり，PDF 形式の論文を表示できるソフトウェアを論文閲覧システムとみなせば，EndNote[4] や Mendeley[5] などの引用文献管理ツールもその 1 つと言える．

書物を読むとき，気になった事柄や，記述に対する疑問を紙面の余白に書き込みながら読書することを Active Reading と呼び，深い読書のためには重要な機能である [1]．この概念を取り入れたシステムとして，タブレット上にペンを使って書き込みがおこなえるデバイス XLibris[4] が 1998 年に発表されている．論文の閲覧に特化したシステムとして，画面上で本文テキストを自由に切り取り再構成できる LiquidText[5] や，本文テキストの行間を引き伸ばしてそのスペースに書き込みをおこなう TextTearing[2] などがある．

一方，内容理解を目的とする論文読解支援システムとして，鉢木らは，論文と Web を連携させ，論文のタイトルページから専門用語を抽出し，その用語を解説するページを論文横に表示するシステムを提案している [10]．石戸谷らは論文と映像の部分要素に対するアノテーションの獲得と蓄積をおこなうための仕組みを開発し，論文と映像を自由に切り替えて閲覧するシステムを提案した [8]．

本システムの特徴は，以上のような論文閲覧と論文読解に必要とされる機能を部分的に取り入れ，さらにサイドノート部分に様々な情報が提示でき，それらを論文レイアウト上で本文と結び付けられるところにある．

3 基本設計

最初に本システムのスクリーンショットを図 1 に掲載する．システムは PC もしくはモバイルデバイスの Web ブラウザ上で動作し，マウス，キーボード，タッチ機能で操作する．論文本文は，3.1.2 節で説明する画像変換から得られた画像そのものを画面中央に表示する．ハイライト部分や特定の領域を画像の上にオーバーレイする形で塗りつぶすことができる．そして本文画像の両サイドには，論文読解を支援する各種リソースを掲載するスペースがあり，必要に応じて，中央の論文本文と線分で接続することが可能である．

[4] http://endnote.com/
[5] http://www.mendeley.com/

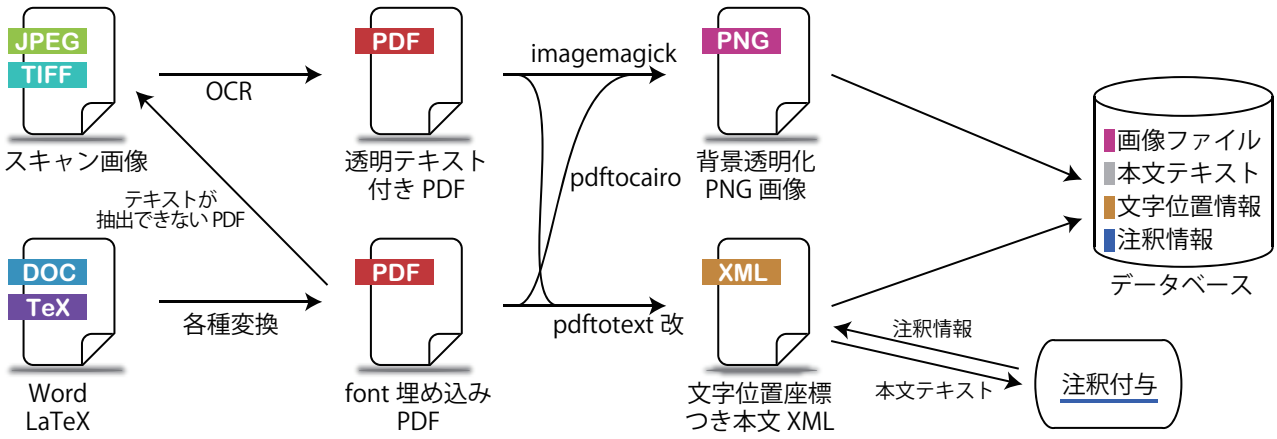


図 2: データ処理の流れ

3.1 データ処理

図 2 に本システムでのデータ処理についてのフローを掲載する．

3.1.1 テキスト情報の抽出

現在，本システムで対応する論文の形式は，紙に印刷された論文および PDF で流通する論文である．紙の論文に対しては，スキャナーを使用してデジタル画像データに変換後，OCR ソフトウェアを用いて文字認識を実行し，透明テキスト付き PDF に変換する．一方，LaTeX や Microsoft Word から作成された PDF 形式の論文はほとんどが PDF 内にテキスト情報が保存されているが，一部の PDF ではフォントのグリフ情報と既存の文字コードの対応表が存在しないものもある．そのような PDF では正しくテキスト情報が抽出できないため，一度画像データに変換した後，OCR を施しテキスト情報を得る．

PDF からテキストを抽出するにあたり，本システムで要求する機能を実現するためには，日本語のような分かち書きのない言語の文字は文字単位で，分かち書きのある言語では単語単位でページ内座標を得る必要がある．しかし，オープンソースソフトウェアライセンスで提供されるプログラムを各種検討したが，条件を満たす使い勝手のよいツールが見つからなかったため，Poppler パッケージ [6] に含まれる pdftotext に独自にパッチをあて対応している．

3.1.2 画像変換

本システムの本文表示で使用している画像形式は，本文ページの背景色を変更する機能に対応するため，透過 PNG 形式を採用している．PDF から PNG への画像変換には，フォント情報を持つ PDF からダイレクトに透過 PNG ファイルを作成できる Poppler パッケージ中の pdftocairo を使用している．スキャン画像から作成された PDF については，imagemagick[7] を用いて背景色 (白色) を透明色に指定した PNG に変換する．

通信トラフィックを減少させるため，Web ブラウザの表示ウィンドウの大きさに合わせて解像度を変えた画像を用意し，白黒のページについては 6 色に，カラーのページでは 64 色に減色した画像を表示に用いている．

4 論文閲覧機能

本節では，論文閲覧に必要であると考える機能を考え，システムの持つ機能を従来の形式 (紙，PDF，XML・XHTML・EPUB)) と比較しながら説明する．

- **本文検索**
  ユーザが指定した任意のキーワードで本文テキストの検索ができる．紙に印刷された論文ではもちろんできないが，PDF や XML 形式では可能である．

- **専門用語の Web 検索**
  5 節で説明する自動用語認識により，特定の専門用語ではクリックすると，その用語の説明ページを表示するか，あるいは Web で検索するかを選択できる．PDF や XML 形式では，ユーザが用語を自ら選択しコピー後，別ウィンドウでテキストボックスにクエリーとして貼り付け検索を実行しなければならない．

- **連続ページめくり**
  紙の論文では，通常，冊子体で出版され，複数の

[6] http://poppler.freedesktop.org/

[7] http://imagemagick.org/

論文をページ順に順番に閲覧できる．一方でPDFやXML形式などの電子媒体では，閲覧単位が1論文なので，次の論文を閲覧する際には，一度リストページに戻る必要がある．本システムでは複数の論文が掲載ページや検索結果など順序付きリストとして定義されているとき，カーソルキーあるいは画面上のボタンを押すことで連続して次々と論文を閲覧できる仕組みを持つ．これにより，一覧ページと論文とを行きつ戻りつする手間を削減できる．

- **拡大縮小表示**
  Webブラウザの画面サイズやモバイルデバイスの画面解像度に応じて，適切な解像度の本文画像を表示とともに，画像の拡大表示の他，2〜4ページまでの割付表示が可能である．紙の論文では文字が小さく，もっぱら電子媒体で論文を読む人には必須の機能である．

- **コントラスト変更**
  紙に印刷された論文は白背景に黒い文字であることが一般的であるが，バックライトが視覚に入るコンピューターディスプレイで，白と黒の組み合わせのハイコントラストを持つ画面を長時間見続けると目が疲れてしまう．そのため背景色を薄い黄や青といった色に変更し，コントラスト値を下げる機能を実装した．

- **柔軟なレイアウト**
  モバイルデバイスでは画面が小さいという制約があるが，紙やPDFのような固定レイアウトのフォーマットでは，2カラムを1カラムで表示するといった画面の大きさに合わせた柔軟なレイアウトの変更ができない．一方でXMLのようなリフロー形式では自由にレイアウトの変更ができる．本システムでは画像ベースの表示のため残念ながらレイアウトの変更はできない[8]．

- **別ページ図表の閲覧**
  論文中の図表はその大きさから，図表の説明文と同一ページに配置できない場合があり，説明文があるページと図表があるページを交互に見ながら論文を読む状況が発生する．リフロー形式では要素の配置を変更すれば同一画面に表示できるが，レイアウト固定フォーマットでは難しい．本システムでは，PDF形式から図表を個別に切り出す処理を開発することで，説明文とともに図表を表示する機能を搭載する予定である．

[8]k2pdfopt: http://www.willus.com/k2pdfopt/ のようなツールを用いれば，固定型のフォーマットでもデバイスの幅に応じたレイアウトの変更ができる

表 1: 論文フォーマットごとの機能特性

	紙	PDF	XML	本システム
本文検索	×	○	○	○
専門用語の Web 検索	手入力	Copy	Copy	Click
連続ページめくり	○	×	×	○
拡大縮小表示	×	○	○	○
コントラスト調整	×	×	○	○
柔軟なレイアウト	×	×	○	×
別ページ図表の閲覧	×	×	○	予定
参考文献リンク	×	×	×	○
書き込み	○	○	○	○
書き込みの集約	×	×	×	予定

- **参考文献リンク**
  本システムでは，論文末尾の参考文献部分から参考文献のリストを取得し，i-Linkageシステム[9]を用いて書誌データベースと同定処理をおこなう．インターネット上で公開されいている文献の場合にはそのリンクを付与し，なければ，論文はCiNii[9]への，書籍はWebcat Plus[10]の該当書誌へのリンクを生成する．

- **書き込み**
  紙の論文では自由に書き込みや下線を引くことができ，PDF形式にも注釈機能がある．本システムにおいても，本文画面上でマウスを用いて領域選択をおこなうと，その領域内のテキスト部分がハイライトされ，コメントの記述やハイライトの保存ができる．

- **書き込みの集約表示**
  紙やPDF形式への書き込みは紙面そのものあるいは論文ファイル内に保存されるため，書き込みの内容と論文を切り離すことが出来ない．そのため，関連文献すべての書き込みを後からまとめて参照することが難しい．本システムでは，書き込みは論文ファイルと分離して保存されるので，ユーザによる書き込みの集約表示が可能である(実装予定)．

最後に，本節で説明した本システムの各機能について従来の論文フォーマットとの比較表を表1に掲載する．

5 論文読解支援機能

本システムは，論文の表示・操作・書き込みといった閲覧に必要な機能とともに，現在表示している論文

[9]http://ci.nii.ac.jp/
[10]http://webcatplus.nii.ac.jp/

のページの本文を解析し，ページの補足情報をページ画像左右の脚注部 (Side-note) に表示する論文読解支援機能を有している．現在表示できる補足情報には次の 2 種類がある．

- **本文中のキーワードに関する情報**
  辞書や百科事典のような見出し語集合とその説明項目というリソースが存在するとき，論文本文中から見出し語を自動抽出し，説明部分を脚注部に表示する．本システムでは，Wikify[3] や Amazon Kindle の X-Ray[11] の技術と同様に Wikipedia をリソースとして用いており，本文中に Wikipedia のタイトル文字列が出現したとき，その説明文と画像を表示している．本文中でマッチしたキーワードは，論文画像の上にオーバーレイでハイライト表示する．現状，キーワードの語義曖昧性解消や表示する説明のランキングなどの精度は高くなく，今後の課題となっている．

  キーワードの説明として，辞書の語釈文のようなあらかじめ静的に定義された文章のほかに，キーワードをインターネットの検索エンジンに渡し，その検索結果を表示する動的な情報提示も可能である．本システムではキーワードを動画検索サイトで検索し，上位 1 位の結果を表示することができる [6]．最近では大学の講義などがインターネット上で数多く公開されているため，キーワードを解説する講義映像がヒットすることが増えてきた．

- **ページの一部と関連する情報**
  表示しているページの本文全部あるいは指定する部分文章に対して，関連する情報を検索し，ヒットした項目を脚注部に列挙する．本システムでは検索アルゴリズムには連想検索エンジン GETA[7] を用いており，検索対象は文書集合として定義できるものならなんでもよく，現在公開されているシステムでは Wikipedia 全ページと，言語処理学会の予稿集を用意している．

これらの情報は，日々刻々と内容が変化するため，Web ページに表示する際に，本文テキストからリアルタイムでキーワード抽出，説明部分の検索，生成がおこなわれる．

6 注釈情報をシームレスに表示する機能

ユーザによる論文本文に対する書き込みやハイライト付与は，本文ページがレイアウトされた状態での注釈付与であり，該当する本文テキストや座標との対応付けを保存しておけばよい．一方，人手による大量の注釈付与作業や，機械的な注釈付与を考えた場合，表示レイアウトそのままでの扱いは難しいため，通常，論文のソースファイルあるいは，テキストフォーマットに変換したデータに対して注釈付与をおこなう．

近年ではソース文書の保存形式として XML フォーマットを採用し，注釈タグを XML 文書に挿入する形式が主流となっているが，付与された注釈情報を PDF のような人間にとって可読性の高いレイアウトに適切な形で表示する仕組みは今のところ存在しなかった．

そこで，我々は注釈情報をレイアウトされた論文画面上に表示するにあたり，XML 中の注釈情報が PDF 中の論文本文のどのテキストと対応するかを求める仕組みを開発した．あらかじめ PDF から座標情報付きのテキストを抽出しておき，XML 中の注釈を指定するタグが囲むテキストと照合し，注釈情報を表すテキストに対し，ページ番号と座標を記録するものである．

本システムでは，論文本文を画像として表示し，本文テキストの座標情報を内部で有しているため，注釈情報を論文画像の上に直接オーバーレイして表示することが可能である．

本機能を利用した例として，論文本文とその発表スライドの対応を示す．最初に論文 PDF から本文のみの XML ファイルを作成し，発表スライドの各ページに対応する論文のテキスト部分を XML タグで囲んだ．次に XML ファイルと PDF ファイルを本システム処理し，表示させた例が図 3 である．本機能により，XML 文書中の注釈情報および，係り受け解析や名詞句のチャンキングなどテキストに対する様々な解析結果を，可読性の高いレイアウトで可視化することが可能になる．

7 おわりに

本稿では，紙に印刷された論文および PDF 形式で流通する論文を対象に，効率的な論文閲覧を実現する機能を有し，論文の本文情報から得られる様々な情報を脚注部に表示する論文閲覧システムを説明した．さらに，論文から抽出したテキストに対し付与した注釈情報を，整形された論文レイアウト上に同時に表示する仕組みを開発した．

現状のシステムでは，論文の内部構造解析として参考文献部分の抽出をおこなっているが，今後，図表領域，およびタイトルやセクションなどの構成要素の認識をおこなう予定である．論文のレイアウト画像上に情報を表示できる本システムは，情報提示法の自由度が高く，これらの認識した要素をどのような形で表示すれば，より効果的に論文閲覧，論文理解になるかを考えていきたい．

[11] http://www.amazon.com/gp/help/customer/display.html/?nodeId=200729910

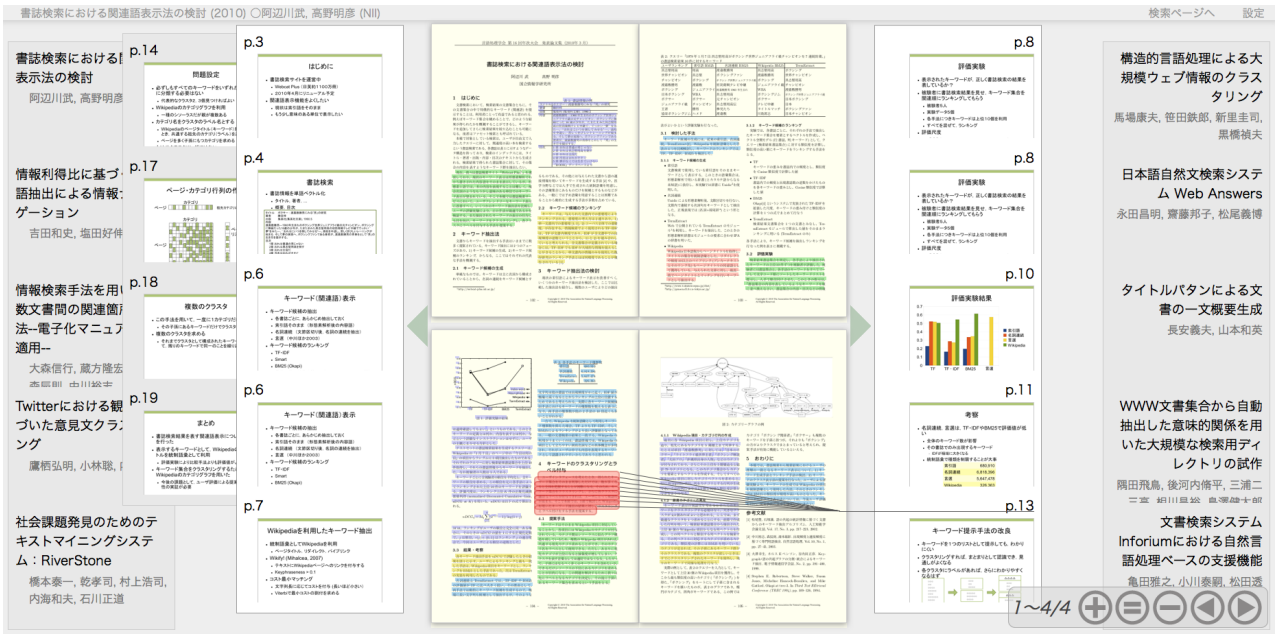


図 3: スライドと論文対応の例

また，大量の論文に対し内部構造解析をおこない，それぞれ認識された要素を有機的に結びつけ，本システムで表示したとき，どのような世界が見えてくるのであろうか? プロトタイプシステムが完成した際には追って報告したい．

参考文献

[1] Mortimer Jerome Adler. *How to Read a Book*. Simon and Schuster, 1940. 邦題:本を読む本. 外山滋比古, 槇未知子訳. 講談社学術文庫. 1987.

[2] Franois Guimbretire Dongwook Yoon, Nicholas Chen. Texttearing: opening white space for digital ink annotation. In *the 26th annual ACM symposium on User interface software and technology*, pages 107–112, 2013.

[3] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *The 18th ACM Conference on Information and Knowledge Management*, pages 233–242, 2007.

[4] Morgan N. Price, Bill N. Schilit, and Gene Golovchinsky. Xlibris: the active reading machine. In *CHI '98 Cconference Summary on Human Factors in Computing Systems*, pages 22–23, 1998.

[5] Craig S. Tashman and W. Keith Edwards. Liquidtext: A flexible, multitouch environment to support active reading. In *CHI '11 Conference on Human Factors in Computing Systems*, pages 3285–3204, 2011.

[6] 阿辺川武 and 間下亜紀子. 文章中のコンテクストに適合した関連動画の検索. In **情報処理学会研究報告エンタテインメントコンピューティング** *2013-EC-27(19)*, 2013.

[7] 西岡真吾. 汎用連想計算エンジン GETA. **コンピュータソフトウェア**, 26(4):87–106, 2009.

[8] 石戸谷顕太朗, 山本圭介, 大平茂輝, and 長尾確. 映像と論文へのアノテーションに基づく論文読解支援システム. **映像情報メディア学会誌**, 66(11):J461–J470, 2012.

[9] 相澤彰子, 高久雅生, and 大山敬三. 大規模データベースを利用したリンケージシステムの提案と実装. *DBSJ Letters*, 6(4):17–20, 2008.

[10] 鉢木稔浩, 太田学, and 高須淳宏. Web 資源を利用した学術論文閲覧支援システム. In **情報処理学会研究報告データベース・システム研究会報告** *2009-DBS-149(14)*, pages 1–6, 2009.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-04

劣モジュラ最適化としての文章情報要約

Document Summarization via Submodular Optimization

河原吉伸 (Yoshinobu Kawahara)[1*]

[1] 大阪大学 (Osaka University)

Abstract: Many criteria for document summarization is known to be submodular functions, which are the discrete counterpart of convex functions. In this paper, we review the recent studies on document summarization based on submodular set-function optimization. And, we also describe some prospects related to this field.

1 はじめに

文章情報要約は，文章を構成する文（または単語）の全体から，要約に用いられる（一部の）文（または単語）を選択する問題であり，本質的に組合せ最適化問題である．この問題は従来からも，整数計画問題や最大被覆問題などの組合せ最適化として定式化され議論されてきた経緯がある [1, 2]．そして特に近年，これらを含む従来から知られる多くの文章要約の基準が，集合関数における凸関数として知られる劣モジュラ関数である事が指摘されている [6]．この事実に基づき，この組合せ的な凸構造に基づく理論的保証を持つ効率的なアルゴリズムの適用や，種々の問題依存の構造を利用した枠組みが可能となる事が報告されている [4, 5]．

本稿では，このような背景から，劣モジュラ性を用いた文章要約に関する最近の動向について概観する．さらに，これを各種の機械学習のタスクにおいて利用する方法について述べる．本稿の以下の構成は，次のようである．まず 2 では，文章要約問題の集合関数最適化としての定式化について述べる．次に 3 では，この際の評価関数における劣モジュラ性との関係について述べ，その重要性について説明する．

2 集合関数最適化としての定式化

まず本節では，文章要約問題の集合関数最適化としての定式化について述べる．なお集合関数 f は，有限の集合 ($\mathcal{V}$ とする) が与えられたとき，その各部分集合 $\mathcal{S}$ ($\subseteq \mathcal{V}$) へ実数を割り当てる関数，つまり $f : 2^{\mathcal{V}} \to \mathbb{R}$ として定義される．

要約の対象となる文章に含まれる文 (sentence) の全体を $\mathcal{V} = \{1, \ldots, N\}$ とする (N は文の数)．このとき

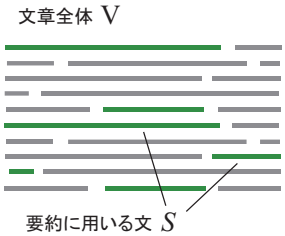


図 1: 文章要約における有限集合 $\mathcal{V}$ とその部分集合 $\mathcal{S}$ の定義の概念図．

文章要約は，要約の質をはかる基準を f を最大化にするような文の集合 $\mathcal{S} \subseteq \mathcal{V}$ を選択する問題として定式化される．このとき，要約の長さに制限があるのが一般的であるため，各文 $i(\in \mathcal{V})$ を選択する事のコストを c_i とすると，次の最適化問題が得られる．

$$\max_{\mathcal{S} \subseteq \mathcal{V}} \ f(\mathcal{S}) \quad \text{s.t.} \ \sum_{i \in \mathcal{S}} c_i \leq b \tag{1}$$

ただし，b は許容される要約の最大の長さである．

要約の質をはかる基準 f としては様々なものが提案されているが，最も一般的なものとしては，次式のように定義される被覆関数

$$\mathcal{C}_i(\mathcal{S}) = \sum_{j \in \mathcal{S}} w_{ij}$$

の和 $\sum_{i \in \mathcal{V}} \mathcal{C}_i$ が挙げられる [2]．ただし，w_{ij} は 2 つの文 i と j の類似性を表す量である．つまり $\mathcal{C}_i$ は，文 i の内容が，選択した文の集合 $\mathcal{S}$ によりどれだけ表されているかという基準になっている．これを文章全体 $\mathcal{V}$ に関して足したものは，選択した文の集合 $\mathcal{S}$ がどれだけ文章全体の内容を表すかを表す基準となる．一般には，要約 $\mathcal{S}$ が十分に文章全体を表されている場合を考慮して，次式のような基準を用いる事が多い．

$$\mathcal{L}(\mathcal{S}) = \sum_{i \in \mathcal{V}} \min\{\mathcal{C}_i(\mathcal{S}), \alpha \mathcal{C}_i(\mathcal{V})\} \tag{2}$$

*連絡先：大阪大学産業科学研究所
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
E-mail: ykawahara@sanken.osaka-u.ac.jp

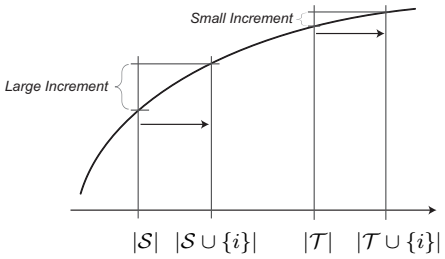


図 2: 劣モジュラ関数の定義 (3) の概念図.

これらの基準は，あとで見る劣モジュラ性 (及び，単調性) と呼ばれる性質を満たしており，効率的に良い解を得られるアルゴリズムを適用する事が可能となる．

3 劣モジュラ性の利用

劣モジュラ性は，集合関数における凸性にあたる離散構造である．上述のように，文書要約は集合関数の最適化として定式化される．その際その基準となる集合関数は，劣モジュラ性を満たす場合が多く，これにより効率的に良い解を得る事ができる．ここでは，そのようないくつかの例について述べる．

なお，人工知能分野における劣モジュラ性の利用に関しては，著者による解説 [9] なども参照されたい．

3.1 劣モジュラ関数とその最大化

劣モジュラ関数は，集合関数における凸性にあたる離散構造であり，1980 年代頃に Lovász により知られるようになった [7]．連続関数における凸性と同様，最小化が効率的に可能であり，局所最適性と大域最適性の一致性や，双対性など凸関数と類似した概念を定義する事ができる．劣モジュラ性には等価な複数の定義が存在するが，次式が直感的にも分かりやすくよく用いられる．

$$f(\mathcal{S}+i)-f(\mathcal{S}) \leq f(\mathcal{T}+i)-f(\mathcal{T}) \qquad (3)$$

ただし，$\forall \mathcal{S} \subset \mathcal{T} \subseteq \mathcal{V}, \forall i \in \mathcal{V} \setminus \mathcal{T}$ である．つまり包含関係にある 2 つの集合 $\mathcal{S}$ と $\mathcal{T}$ に関して，包含される集合 $\mathcal{S}$ へ新しい要素 i を加えた際の増分が，包含する集合 $\mathcal{T}$ の場合のそれより大きくなる (図 2 参照)．このように劣モジュラ関数は，サイズと共に増加が穏やかになる性質を持っており，限界効用逓減の法則を表す関数としても知られる．また，任意の $\mathcal{S} \subseteq \mathcal{T} (\subseteq \mathcal{V})$ に対して，集合関数 f が $f(\mathcal{S}) \leq f(\mathcal{T})$ を満たすとき，f は単調非減少であると言う．

なお上述の被覆関数は (単調非減少) 劣モジュラ関数であり，その他の基準も劣モジュラ性を満たす場合が多い [6]．これは，多くの文を使うほど，元の文章全体の内容を表せる傾向が高くなる事からも直感的に理解できる (図 3 参照).

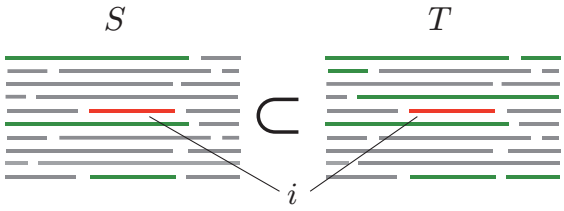


図 3: 文章要約における評価関数の劣モジュラ性 (3) の概念図.

Algorithm 1 どん欲法の手順

1: $\mathcal{S}_0 \leftarrow \emptyset, i = 1$ に設定.
2: **while** $i \leq k$ **do**
3: $\quad \rho^+_{j_i}(\mathcal{S}_{i-1}) = \arg\max_{j \in \mathcal{V} \setminus \mathcal{S}_{i-1}} \rho^+_j(\mathcal{S}_{i-1})$ となる要素 $j_i \in \mathcal{V} \setminus \mathcal{S}_{i-1}$ を選択.
4: $\quad \mathcal{S}_i \leftarrow \mathcal{S}_{i-1} \cup \{j_i\}, i \leftarrow i+1$.
5: **end while**

上述のように，文章要約によく用いられる被覆関数は最大化される事で，文章を要約する文の集合 $\mathcal{S}$ を選択する．このように文章要約は，(単調非減少) 劣モジュラ関数の最大化として定式化される事が多い (つまり，式 (1) における f が単調非減少劣モジュラ関数)．一般に，(単調非減少) 劣モジュラ関数の (サイズ制約下での) 最大化問題は NP 困難な問題であるが，どん欲法と呼ばれる単純なアルゴリズムにより，理論的に，かつ実用的に良い近似解が得られる事が知られている [8]．どん欲法の手順は，Algorithm 1 に示すように単純なものであるが，最悪ケースでも最適解の $(1-1/e) \approx 0.632$ 倍の値を持つ近似解を与える事が知られている (e は自然対数の底)．ただし $\rho^+_j(\mathcal{S}) := f(\mathcal{S} \cup \{j\}) - f(\mathcal{S}))$ である．経験的にも，多くの場合で貪欲法により極めて良い解が得られる事が報告されている [8].

3.2 要約基準と劣モジュラ性

文章情報要約のための基準は，これまで様々なものが提案されてきた．これらは一般に，選択する文の冗長性をできるだけ除外する，というのが基本的な考え方であるものが多いが，劣モジュラ性を満たす事が知られている．

例えば，一般的によく用いられる基準として，次式のように定義される (Maximal) Marginal Relevance と呼ばれる基準がある [1].

$$f(\mathcal{S}) = \sum_{i \in \mathcal{S}} [\lambda(Sim(i,Q) - (1-\lambda) \max_{j \in \mathcal{V}} Sim(i,j))]$$

ただし，$\lambda \in [0,1]$，Sim は何らかの類似尺度，Q はクエリである．この基準も劣モジュラ関数である事が示されている．また別の例としては，先の被覆関数 (2) に加え，各 $\mathcal{S}$ を選択する事に対する利得 $\mathcal{R}(\mathcal{S})$ を定義し，これらを加え合わせた基準も提案されている．

$$f(\mathcal{S}) = \mathcal{L}(\mathcal{S}) + \gamma\mathcal{R}(\mathcal{S})$$

$\mathcal{R}(\mathcal{S})$ としては，次式のように，選択する文が分散する事に対して利得を与えるようなものが知られる [6]．

$$\mathcal{R}(\mathcal{S}) = \sum_{i=1}^{K} \sqrt{\sum_{j \in P_i \cap \mathcal{S}}}$$

ただし，P_i $(i = 1, \ldots, K)$ は $\mathcal{V}$ の分割を表す．また，人による要約との比較に基づいた，ROUGE-N[3] と呼ばれる一般的に用いられる基準も，劣モジュラ性を満たしている事が知られている．

このように，要約に用いられる多くの基準は，劣モジュラ性を満たした集合関数である．従ってその最大化に関しては，劣モジュラ性のために，理論的保証のある近似解がどん欲法により効率的に得られる．また，より実用的なアルゴリズムなども多数提案されており，これらを問題に応じて適用する事で大規模な場合などでも適用可能な自動要約が可能となると言える．

4 むすび

本稿では，劣モジュラ関数最大化としての文章要約に関する定式化について述べた．劣モジュラ関数最大化は，どん欲法により効率的に理論保証のある近似解が得られる事が知られており，その他の実用的なアルゴリズムも多数提案されている．これらを適用する事により，実用的であり，かつ理論的保証のある文章要約を行う事ができる．

本稿ではふれなかったが，劣モジュラ最適化としての定式化をベースにする事により，機械学習で扱われる様々な問題と共通の枠組みの中で議論する事ができるようになる．これにより，文章と，その他のデータ (画像など) とを融合的に用いた数理的枠組みを実現する事も可能であると思われる．

参考文献

[1] J. Carbonell and J. Goldstein. The use of MNR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the Ann. Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'98)*, pages 335–336, 1998.

[2] E. Filatova and V. Hatzivassiloglou. A formal model for information selection in multi-sentence text extraction. In *Proc. of the 20th Int'l Conf. on Computational Linguistics (COLING'04)*.

[3] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, 2004.

[4] H. Lin and J. Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Proc. of the 48th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'10)*, pages 912–920, 2010.

[5] H. Lin and J. Bilmes. Word alignment via submodular maximization over matroids. In *Proc. of the 49th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*, 2011.

[6] H. Lin and J.A. Bilmes. A class of submodular functions for document summa- rization. In *Proc. of the 49th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*, pages 510–520, 2011.

[7] L. Lovász. Submodular functions and convexity. In A. Bachem, M. Grötschel, and B. Korte, editors, *Mathematical Programming – The State of the Art*, pages 235–257. 1983.

[8] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions – I. *Mathematical Programming*, 14:265–249, 1978.

[9] 河原吉伸，永野清仁，鷲尾隆. 劣モジュラ性を用いた知能情報処理への新展開. **人工知能学会誌**, 27(3):252–260, 2012.

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-05

Twitterにおけるトピック遷移分析システムの提案

A Proposal of a Topic Transition Analysis System for Tweets

田中克明 [1*]

[1] 一橋大学情報基盤センター
[1] Center for Information and Communication Technology, Hitotsubashi University

Abstract: In this paper, we propose an interactive system to represent the transition of topics extracted from documents that are generated in chronological order, such as tweets. Many of methods, extracting and visualizing topic transitions in documents generated along the time series aim to show an overview. We implement a system, reorganizing and visualizing topic transitions based on keywords designated by a user, providing interfaces to read the original documents for user to support analyzing topic transitions.

1 はじめに

本研究では，Twitter など時間経過に伴い生成される文書集合にふくまれる時間経過に沿ったトピックの遷移を，インタラクティブに提示する仕組みを提案する．時系列に沿って生成される文書からそこに含まれる内容を抽出し可視化する手法の多くは，表示をユーザが目視することにより全体的な概要を理解することの支援を目的とする．それに対し，本稿で提案するシステムは，トピック遷移をそのまま提示するだけではなく，ユーザが指示した単語などの情報に基づき，トピック遷移の一部分を抽出し提示し，さらにトピックに含まれる文書の詳細をユーザが確認可能とすることにより，ユーザの興味に応じ，時間経過に沿ったトピック遷移の分析を支援するシステムを実装，提案する．

本稿では Twitter から取得したツイートからトピックを抽出するために，Probabilistic Latent Semantics Indexing（pLSI）[Hofmann 99] を用いた．pLSI により，文書からトピック z，文書中にあらわれる単語 w についてトピックごとの生起確率 $p(w|z)$，各ツイート d のトピックにおける生起確率 $p(d|z)$ などを求めることができる．これらの確率を用いて，実装したシステムでは，ユーザが興味を持った単語の生起確率が大きいトピックからなるトピック遷移を表示，そこから個別のトピックを選択し，トピック内での生起確率が高い単語や含まれるツイートの確認を可能とした．これにより，ツイート本文などを確認した後に，新たに興味をひかれた単語やツイートを指定し，それらを含むトピックの推移をあらためて確認するなどインタラクティブな分析が行える．

*連絡先： 一橋大学情報基盤センター
〒 186-8601 東京都国立市中 2-1
E-mail: sigam07@katsuaki-tanaka.net

2 関連研究

2.1 トピック遷移の抽出と可視化

本研究において提案するシステムの分析対象である，時間経過に沿ったトピック遷移の抽出のための手法は，トピックモデルに基づく Dynamic Topic Models[Blei 06] などが挙げられる．これらの手法では，時系列に沿って時区間を設け，その区間に対し一定数のトピックを抽出する．また，k-means を拡張し古い情報を忘却するモデルを取り入れたクラスタリング手法 [長谷川 07] の研究もなされている．

抽出したトピックの可視化は，特徴語を並べる，トピック出現確率の推移をグラフ化するなど以外に，全体の傾向を把握しやすいように可視化を行う，Themeriver [Havre 02] や Alluvial Diagram [Rosvall 10] などが研究されている．可視化結果は静的なものに限らず，一部を選択し強調表示するなどの操作が可能なものもある．

トピックの遷移を操作するためには，遷移をトピックとトピック間のリンクからなるグラフ構造とし，Gephi [Bastian 09] などのグラフ構造可視化ツールを用いる方法が考えられる．これにより，遷移の構造を可視化すると同時に，ノード（トピック）の表示・非表示，グラフ構造の変形などの操作を行うことができる．しかし，トピックに含まれる単語ごとの出現確率に応じた操作など，トピックの抽出過程で得られたデータを活かし，グラフの要素であるノードに対し細かな操作を行うためには，グラフの元となるデータの再生成が必要であり，グラフ可視化ツールは，文書・単語からなるトピック遷移のインタラクティブな操作には不十分である．

2.2 Twitter データの分析

Twitter のデータに対する分析は，ユーザ間の関係に関する研究 [風間 10][Cha 10]，時系列データとして一時的な増大などに着目した研究 [Sakaki 10][水沼 13]，タイムラインからのツイート間の構造抽出に関する研究 [松尾 14] などがなされている．

また，本稿で扱うデータと同様に,「人工知能」を含むツイートに対し，特徴語の推移，ツイートしたユーザに関する分析などが行われている [鳥海 14]．

3 時系列トピック遷移の抽出

提案システムが取り扱う時間の経過に沿ったトピックの遷移をツイート群から以下の手法により抽出した．なお，提案するシステムでは，一定の時区間（区間数 N）ごと各自区間における K 個のトピック $z_{n,k}(n = 1,2,\ldots,N,\ k = 1,2,\ldots,K)$ と，トピックの生起確率 $p(z_{n,k})$，各トピックにおけるツイート d_i の出現確率 $p(d_i|z_{n,k})$，単語 w_m の生起確率 $p(w_m|z_{n,k})$ を利用する．これらを求めるために，筆者が人工衛星の設計議事録からのタスク抽出に用いた手法 [Tanaka 11] を改良してトピック遷移の抽出を行った．

3.1 前処理

トピック抽出の前に，処理対象とするツイートを，Twitter REST API の search/tweets により収集する．同 API で収集できるツイートは過去約 1 週間分に限定され，長期間にわたり収集するために，定期的な API 呼び出しを行った．得られたツイート群からは，タイムラインでのツイートの扱いを模して公式リツイートを除去した．また各ツイートからは，URL，リツイートまたは引用ツイートを示す「RT」「QT」に続くテキストを取り除いた．これらを MeCab[1]を用いて形態素解析し，名詞および未知語と分類された語とその出現回数を求め，各ツイートに対応する単語ベクトル d_i を得た．

3.2 トピックの抽出

処理対象とするツイートのツイートされた時刻に着目し，最も古いものと最も新しいものの間を N の区間に分割，各区間の終了時刻 t_n をもとめる．ここでは，$N = 50$ とした．処理対象とするツイートのうち $t_n(n = 0,1,2,\ldots,N)$ 以下の時刻を持つツイートにより，ツイート集合 D_n を設定し，pLSI により K 個のトピック $z_{n,k}$ を抽出した．ツイート d_i に対して pLSI により求められた $p(z_{n,k})$，$p(d_i|z_{n,k})$ を用いて，

$$\arg\max_k p(d_i, z_{n,k}) = \arg\max_k p(z_{n,k})p(d_i|z_{n,k}).$$

をとる k を持つクラスタ $C_{n,k}$ へ，排他的なクラスタリングをあわせて行った．以後，$z_{n,k}$ あるいは対応する $C_{n,k}$ を，t_n におけるトピックとして扱う．提案システムではトピックの遷移全体の俯瞰ではなく，その中でユーザが着目した部分を扱うことであるため，トピック数 K は大きめにとった．

3.3 古いツイートの忘却

新しいツイートと関連を持たないツイートは，古い内容であり，時間の経過に従い徐々に忘れ去られていくと考えられる．そこで，D_n からトピック抽出を行う前に，古いツイートの重みを徐々に減らす忘却の仕組みをもうけた．

古いトピックとは，トピック $z_{n,k}$ に対し，$p(d_i|z_{n,k})$ が大きいものから順に見ていき $\sum_{i\in C_{n,k}} p(d_i|z_{n,k}) \leq S(S = 0.2)$ の間に存在する d_i において，ツイートされた時刻が $t_(n-1)$ より小さい，すなわち新しいツイートを含まないトピックを指すこととした．一方，ツイートされた時刻が t_n より大きい，すなわち新しいツイートを含めば，$z_{n,k}$ を新しいトピックとみなす．

新しいトピックに含まれない d_i に対し，D_{n+1} からトピック抽出を行う際に $R(R \leq 1)$ を乗じ，古いツイート d_i が徐々に忘れ去られるようにした．

3.4 トピック間遷移の設定

抽出したトピック間の類似度を以下の $sim(C_{n,i}, C_{n+1,j})$ と定義し，表示時に閾値 T 以上の類似度を持つ $C_{n,i}, C_{n+1,j}$ に対し，リンクを設けた．

$$sim(C_{n,i}, C_{n+1,j}) = \frac{|C_{n,i} \cap C_{n+1,j}|}{|C_{n,i}|}. \quad (1)$$

クラスタなど，複数の要素からなる集合の類似度は，次の Jaccard 係数により求めることが多い．

$$Jaccard(C_{n,i}, C_{n+1,j}) = \frac{|C_{n,i} \cap C_{n+1,j}|}{|C_{n,i} \cup C_{n+1,j}|}. \quad (2)$$

ツイートは時間がたつほど数が増えるため，$C_{n,i}$ に比べ $C_{n+1,j}$ の方が要素数が多いと考えれ，$C_{n+1,j}$ の要素数が大きいと Jaccard 係数は小さな値を示し，類似度が低く判定されるため，(1) を類似度として用いる．

[1] http://mecab.sourceforge.net/

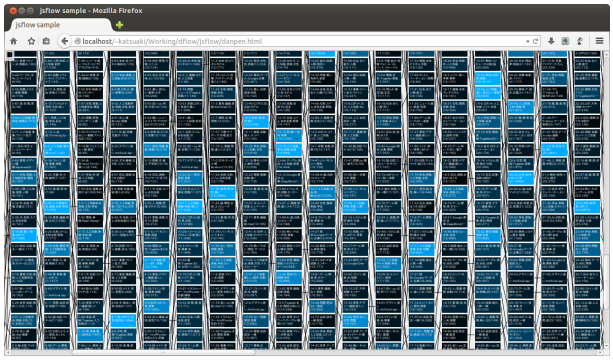


図 1: トピック遷移表示例

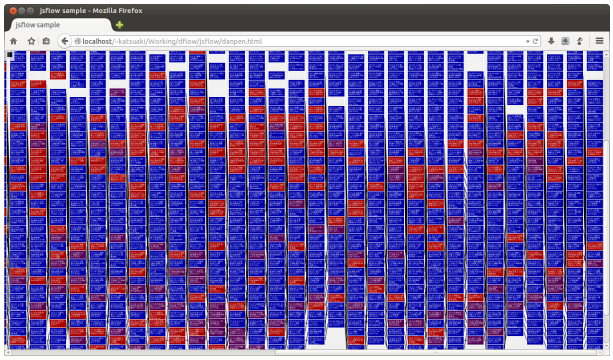


図 2: 「人工知能」(青)「表紙」(赤) を指定した例

3.5 トピック遷移の表示

ここでは，トピックをノード，トピック間の類似度が閾値以上のものをリンクとして得たグラフ構造を，時間を横軸にとり表示した．ラベルには，各トピック $z_{n,k}$ において $p(w_m|z_{n,k})$ が大きい語を選択した．表示例を図 1 に示す．

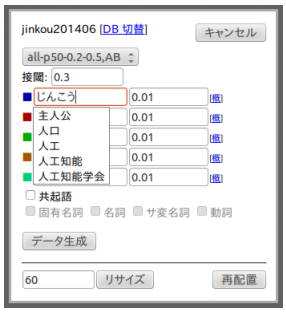
4 システム概要

ここから，本研究で提案するシステムで実装した，トピック遷移分析システムの各機能について述べる．

4.1 単語の生起確率によるトピック遷移の選択

3.5 にて述べたトピックの遷移全体の表示に対し，本システムのユーザがキーワード w と閾値を指定することにより，キーワードの生起確率 $p(w|z_{n,k})$ が閾値以上のトピック $z_{n,k}$ を選択し，指定された色により表示する．すなわち，トピック遷移のうちキーワードに関連する部分を抽出して表示する．

ひとつのキーワードを指定すれば，そのキーワードを含むトピックを，複数のキーワードを指定すれば，各



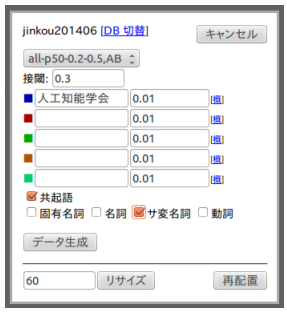


図 3: キーワード入力支援例 図 4: 単語ラベルの指定例

キーワードにまつわるトピックの移り変わりを表示することが可能である．図 2 に例を示す．キーワードの生起確率閾値の設定には，後述する単語出現状況の表示における $p(w|_z n,k)$ の推移が参考になる．

4.2 ラベル語の指定

ラベルとしてキーワードと同じツイートに含まれる単語，すなわち共起する単語を選択することを指定すると同時に，形態素解析時に得られた単語の品詞を指定することができるようにした．画面例を図 4 に示す．

キーワードとして文書群に含まれる何らかの「着目対象」を指定すると，着目対象に対してどのような議論が行われていたかを表示できる．同時に，ラベルとして表示する語の品詞として，サ変名詞 (「〜する」と「する」を続けられる名詞) を指定すると，着目対象に対して行われていた行為を抽出できる．これにより，ある対象への作業の一覧を確認することができる．また，時間経過に沿ったトピック抽出を経ているため，同じタイミングで並行して行われていた事象を分離することが可能である．

4.3 キーワード入力支援

ユーザがキーワードの入力を行う際，キーストロークを含む単語を文書に含まれる単語リストから取得，再構成用のキーワード候補として表示する仕組みを設けた (図 3)．

入力支援を行うことにより，文書中に確実に存在する単語を確実に入力できるようにすることを目指した．一方，キーワード入力支援を行わない場合，ユーザが，表記の揺れなど含まれる単語を把握した上でキーワードを指定する必要が生じる．また，入力支援により，例えば「人工知能」と「人工知能学会」の両方が単語として本システムに認識されている場合，両者を候補として同時にユーザに表示することにより，語の違いを意識してキーワードを指定する必要性を示せる．

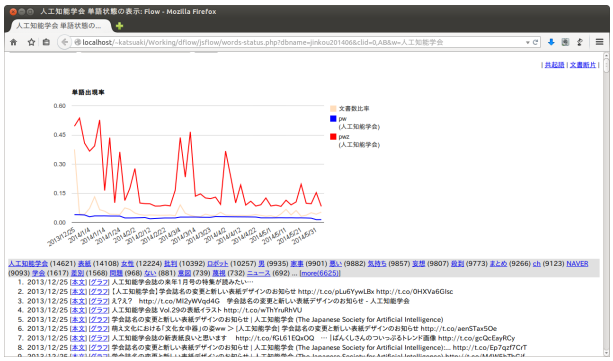


図 5: 単語出現状況の表示例

図 6: トピック詳細の表示例

4.4 単語出現状況の表示

キーワードの指定画面から，キーワードとして設定しトピックの選択を行う前に，キーワード候補である単語のトピック遷移内での出現状況を表示させられるようにした．図 5 に例を示す．単語の出現状況表として表示するのは，単語 w について pLSI により求められた $\max p(w|z_{n,k})$ と $p(w)$ の推移を示すグラフ，ツイート内に共起するその他の単語，単語を含むツイートの一覧である．

本表示における単語の出現確率の遷移を示すグラフは，4.1 に述べたトピックの選択表示のためのキーワードと閾値となる $p(w|z_{n,k})$ を設定する支援となる．また，ツイート内で共起する単語の表示を行うことで，複数の単語を指定する場合の 2 番目以降のキーワードの選択を支援することも目指した．

4.5 トピック詳細の表示

$z_{n,k}$ において，$p(d_i|z_{n,k})$ の大きい d_i 順，あるいはツイートされて時刻が新しい順に，ツイートを表示する．また，$p(w_m|z_{n,k})$ が大きい順に単語 w_m も表示する．図 6 に例を示す．これにより，トピックの詳細を把握することができる．また，各ツイートについて，Twitter

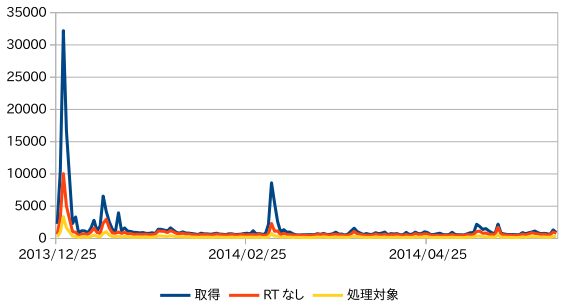


図 7: 取得ツイート・処理対象ツイート数の日次推移

Web クライアント上により実際のツイートを参照できるようにした．

4.6 ツイートを含むトピックの表示

単語の詳細表示画面，トピック詳細の表示画面には，単語を含むツイート，トピックに含まれるツイートの一覧が表示される．ここから，ツイート d_i を指定し，$p(d_i|z_{n,k})$ が大きい上位 100 個の $z_{n,k}$ のトピックを選択，表示する機能をもうけた．これにより，指定したツイートがトピック遷移の中でどの期間にわたって主に出現し，どのようなトピックへ含まれているかを確認できるようにする．

5 「人工知能」を含むツイートにおける利用事例

処理対象の例として,「人工知能」を検索クエリとして Twitter API により収集，2013 年 12 月 25 日 19 時付近からから 2014 年 6 月 6 日 18 時付近（どちらも JST）までの 235,979 ツイートを得た．これらより 3.1 に述べたように公式公式リツイートを除去した 131,522 から，以後の処理では処理量を減らすために，約 $\frac{1}{3}$ にあたる 43,862 ツイートをランダムに選択した．選択されたツイートに 3 以後のトピック遷移抽出処理を行い，以後の事例確認に用いた．Twitter より取得したツイート数と処理対象としたツイート数などの日ごとの推移を図 7 に示す．

5.1 トピック遷移の選択とラベル語指定

「人工知能学会」「表紙」の 2 つの単語を指定してトピックの抽出を行うと，両者が混じり合いながらツイートが続いている様子がわかる．このうち,「表紙」が含まれないトピックの一部を確認すると，人工知能学会

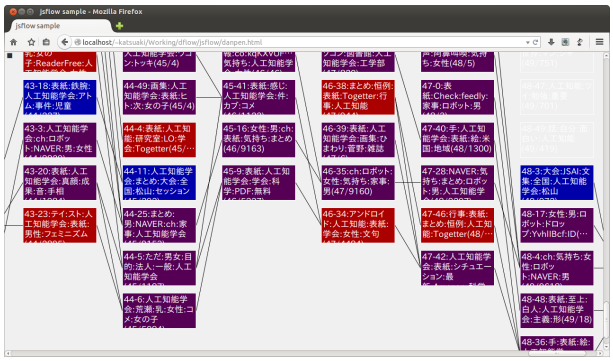


図 8: 「人工知能学系」(青)「表紙」(赤)を指定しラベルを共起する名詞にした例

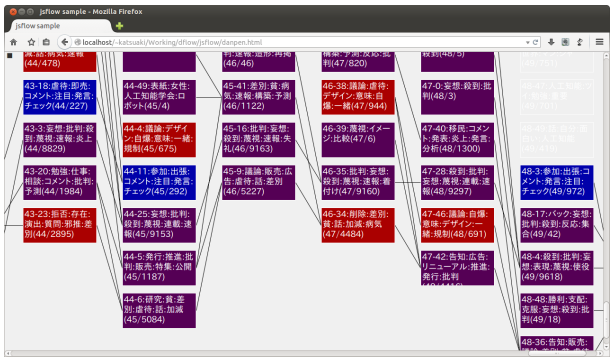


図 9: 「人工知能学会」(青)「表紙」(赤)を指定しラベルを共起するサ変名詞にした例

全国大会について述べているツイートを含むトピックであった．トピックの遷移表示において，ラベル語を共起する名詞とした場合を図 8 に，共起するサ変名詞とした場合を図 9 に示す．名詞をラベルにすると，どのような事象があったかを確認でき，サ変名詞をラベルにすると，どのような意図の記録としてツイートされているかを確認することが，おおよそ可能である．

5.2 トピックとツイートの参照

「人工知能」を含むツイートを分析した研究 [鳥海 14] にて，BBC などにて人工知能学会誌表紙が取り上げられた旨の記述があることから,「BBC」について確認した．はじめに図 3 のキーワード入力画面にて「BBC」を入力しようとしたところ，4.3 のキーワード入力支援により「BBC」が候補として表示され，ツイートに現れ単語として認識されていることがわかった．続いて 4.4 の単語の状況表示より,「BBC」のトピック遷移中での出現確率の推移を確認した．これに基づき，トピック中に「BBC」が出現すると判断する閾値を設定，4.1 のトピック選択を実行する．選択表示されたトピックの詳細を 4.5 のトピック詳細表示により表示することにより,「BBC」を含むトピックに含まれるツイートを確認することができる．この際，図 6 にも示したトピックより,「AFP」も表紙に関わる報道を海外向けに行なっていることがわかった.「BBC」同様に「AFP」について確認を行うと，AFP がいくつかの国にニュース配信を行ったことに触れたツイートを発見できた．

6 考察

本稿では，大量のツイートに対し，トピック抽出により得られたトピックの遷移をユーザの指示するキーワードに基づき提示する機能，それらトピックに含まれるツイートや単語の詳細を確認する機能などを持った分析システムを提案，実装した．本システムでは，ユーザが興味をひかれた事象について，代表的な単語などにより全体を俯瞰するだけではなく，個々のツイートに含まれる内容を読み込むことが可能である．これにより，ユーザがはじめに興味をひかれた事象の詳細を確認するうちに，あらたな事象に興味を持ち，トピックの遷移全体での新たな興味対象の位置づけを確認し詳細を読み込むという行為を繰り返し，ネットサーフィンに似たような形で，トピックの遷移を確認していくことができる．

既存のトピックの遷移抽出や抽出結果の提示手法は，抽出対象とした文書集合全体におけるトピック遷移の位置づけの提示を主な目的としている．また，対象として，報道記事や論文を扱っており，結果を見る側が処理され提示される文書集合の内容に対し，ある程度の知識を持っていることが暗黙の前提になっていると考えられる．例えば，今回取り扱った「人工知能」を含むツイートにおいて，学会誌の表紙について議論が起きたことを知っているため,「家事」「批判」などの特徴語の表記で何が議論されているかわかが、知らなければ人工知能と「家事を批判すること？」の関係は類推しづらい．

一方，本稿で提案するシステムでは，複数のトピックが提示され，含まれるツイートをひとつずつ確認することが可能であり，興味を持った部分から詳細を読み進めることにより，内容に関する前提知識がなくても，理解できる文から読みはじめることができる．

このように，Twitter 上の情報の理解を支援することが可能ではあるが，研究を進めるためは，どのような内容についてどの程度の支援が可能であるか，評価を行なう必要がある．

また，分析対象としたツイートを確認すると，それぞれある事象について感想などの形で言及が多く，現実世界のコピーとしての情報が多い．そのため，時間経過に沿って抽出されたトピックを確認すると，新しく雑誌が発行されたなど起こった事象は反映されてい

るが，Twitter 内部で時間の経過に沿って進んだ議論を見つけることができない．Twitter 自体のもつ性質にも依存するが，Twitter 上で，言及のあとどうするかという「議論」が起きていないわけではない．例えば，キーワードを含むリツイートに続いて投稿されたツイートは，リツイート内容についての意見を含んでいる可能性がある．キーワードを指定した検索では，言及以外のツイートを収集することが出来ないため，検索結果の前後のツイート，リプライ関係にあるツイートなども含めて収集し，分析対象とするか判断する必要があろう．

現在の提案システムの実装では，どのトピックを参照したかなどの履歴が残らないため，意図せず繰り返し同じトピックを参照してしまうなど，操作上の不都合がある．トピックやツイートにブックマークをつける，メモを付記するなどの機能，キーワードの生起確率によりトピックを選択する際にキーワードを含まないトピックを選択する機能，トピック選択時のデータ生成速度，トピックを示すグラフ上のノードの色付け方法など，改善すべき点が多い．

7 おわりに

本稿では，長期間の大量のツイートに対し，そこに含まれるトピックの遷移をユーザの興味に基づいて表示しつつ，ツイート本文までユーザが読み進めることができる仕組みを実装した．提案したシステムにより，概要を眺めるのでもツイートをすべて読むのでもなく，ツイートの拾い読みを支援するような形で，面白そうな部分を渡り歩くことが可能である．今後，分析対象とするデータと分析システムの整備を行いつつ，提案システムによる分析でどのような事柄の理解が可能か，評価を進めたい．

参考文献

[Bastian 09] Bastian, M., Heymann, S., and Jacomy, M.: Gephi: an Open Source Software for Exploring and Manipulating Networks, in *Proceedings of Third International AAAI Conference on Weblogs and Social Media*, pp. 361–362 (2009)

[Blei 06] Blei, D. M. and Lafferty, J. D.: Dynamic Topic Models, in *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120 (2006)

[Cha 10] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K.: Measuring User Influence in Twitter: The Million Follower Fallacy, in *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media*, pp. 10–17 (2010)

[Havre 02] Havre, S., Hetzler, E., Whitney, P., and Nowell, L.: Themeriver: Visualizing Thematic Changes in Large Document Collections, *Visualization and Computer Graphics, IEEE Transactions on*, Vol. 8, No. 1, pp. 9–20 (2002)

[Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Indexing, in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57 (1999)

[Rosvall 10] Rosvall, M. and Bergstrom, C. T.: Mapping Change in Large Networks, *PloS one*, Vol. 5, No. 1, p. e8694 (2010)

[Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, in *Proceedings of the 19th international conference on World wide web*, pp. 851–860 (2010)

[Tanaka 11] Tanaka, K. and Hori, K.: Extracting Tasks in Design Process Records, in *Proceedings of Eighth International Joint Conference on Computer Science and Software Engineering*, pp. 373–378 (2011)

[松尾 14] 松尾 哉太, 新妻 弘崇, 太田 学：Twitter タイムラインの話題の可視化の一手法, 第 6 回データ工学と情報マネジメントに関するフォーラム (2014)

[水沼 13] 水沼 友宏, 池内 淳, 山本 修平, 山口 裕太郎, 佐藤 哲司, 島田 諭：Twitter におけるバーストの生起要因と類型化に関する分析, 情報社会学会誌, Vol. 7, No. 2, pp. 41–50 (2013)

[長谷川 07] 長谷川 幹根, 石川 佳治：T-Scroll: 時系列文書のクラスタリングに基づくトレンド可視化システム, 情報処理学会論文誌, Vol. 48, pp. 61–78 (2007)

[鳥海 14] 鳥海 不二夫, 榊 剛史, 岡崎 直観：「人工知能」の表紙に関する Tweet の分析 (小特集「人工知能」表紙問題における議論と論点の整理), 人工知能：人工知能学会誌：journal of the Japanese Society for Artificial Intelligence, Vol. 29, No. 2, pp. 172–181 (2014)

[風間 10] 風間 一洋, 今田 美幸, 柏木 啓一郎：Twitter の情報伝播ネットワークの分析, 第 24 回人工知能学会全国大会 (2010)

人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-06

コミックを対象とした質問応答システムのための質問タイプ分類の検討

Qestion Type Classification for Comic QA System

山下 諒[1] 陸 鑫一[2] 松下 光範[3*]
Ryo Yamashita[1] Xin-yi Lu[2] Mitsunori Matsushita[3]

[1] 関西大学大学院 総合情報学研究科
[1] Graduate School of Informatics, Kansai University
[2] アズワン株式会社
[2] As One Corporation
[3] 関西大学 総合情報学部
[3] Faculty of Informatics, Kansai University

Abstract:
The objective of our research is to realize a question answering (QA) system for comics. Because comic is a multi-modal contents that utilizes texts and illustrations cooperatively, question sentence that should be handled by the comic-QA system varies significantly in comparison with the conventional QA system. To meet this goal, this paper performs type classification of the question for comics as a basic examination. We classified question sentences into query types: bibliographic information type questions (5 types) and content information type questions (6 types). These types are determined by the result of previous works and question sentences collected from Web sites. We performed automatic classification based on the classification. As the result, we observed that accuracy was high in bibliographic information type question,while that in content information type question was low.

1 はじめに

近年，スマートフォンやタブレットなどの電子端末が急速に普及しつつある．それに伴い，コミックのディジタル化が急激に進んでいる．

コミックがディジタル化されることによって，紙媒体のコミックが有していた物理的な制約から開放され，従来のコミックの枠にとらわれない表現 (e.g., 話の展開に応じて内容を切り替える，コマに動きを付与する) が可能になるだけでなく，コミックの書誌情報だけでなくコンテンツそのものに対する柔軟な情報アクセス (e.g., 読み手の母語に応じて言語を切り替える，特定のキャラクターが出現するページを検索する) も可能になると期待される．現状の電子コミックの多くは，紙媒体のコミックをページごとの画像情報として電子化したものが大半であるが，それらを対象として，コマの同定 [2, 12] や登場キャラクターの特定・抽出 [7, 13]，書誌情報やコンテンツ情報の構造化 [1] などの研究が精力的に推められており，近い将来には，コミックに対する，より柔軟な情報アクセスが可能になると考えられる [4].

本研究では，こうした柔軟な情報アクセスが可能な状況を前提として，コミックを対象とした質問応答の実現を目指している．質問応答は現在自然言語処理分野でテキストを対象として精力的に進められている研究の一つであり [5]，コミック質問応答はそれをコミックコンテンツに拡張したものである．

コミックは，テキストやイラストなどの複数の要素から構成されたマルチモーダルなコンテンツである．そのため，テキストを対象としている場合に比べて問題は飛躍的に難しくなる．例えば質問応答の場合，「ドラえもんで『もしもボックス』が初めて登場したのは何巻ですか?」という質問であれば，「小学館てんとう虫コミックス 11 巻です」とテキストで応答するのが適切であるが，「スラムダンクで『あきらめたらそこで試合終了ですよ』と安西先生が言ったコマが見たい」や

*連絡先：関西大学大学院総合情報学研究科
〒 569-1095 大阪府高槻市霊仙寺町 2-1-1
E-mail:mat@res.kutc.kansai-u.ac.jp

「名探偵コナンで主人公がスケートボードに乗っているシーンが見たい」といった質問の場合は，コミックのコマやストーリーの一部分を提示するのが適切であろう．このように，コミック質問応答を実現するためには，様々なユーザの質問に応じて，回答を生成する戦略を切り替えたり，コミックコンテンツの中から応答として適切な箇所を同定したりする技術が必要になる．

本稿では，こうしたコミック質問応答の要素技術の 1 つである質問文理解技術に着目し，その実現のためにコミックを対象としたコミックに関する質問を収集し，それらの質問タイプの分類を試みる．

2 関連研究

2.1 質問応答技術

質問応答技術とは，一般的な検索エンジンとは異なり，質問に対して直接回答を提示する技術を指す．質問応答の研究の歴史は古く，構造化されたデータベースを対象に，自然文で表現された質問を通じて条件を満たすデータを検索する技術の研究 [3] が 1960 年代から行われているが，近年の自然言語処理分野で盛んに研究されている質問応答は，Web などから得られるテキストデータ集合を対象として，質問に合致する情報を検索・抽出して提示する技術である [5]．以下では，後者の質問応答を Web 質問応答と記す．本研究で実現を目指すコミック質問応答は，この Web 質問応答の枠組みを延伸したものである．池野らは，一般的な Web 質問応答システムの基本的なプロセスを質問解析，情報検索，情報抽出，回答選択と整理している [10]．コミック質問応答システムの場合，Web 質問応答システムとは異なり，上述したようなマルチモーダルなコンテンツを対象とするため，必ずしもこのプロセスに当てはまるわけではない．そこで次節では，コミックを対象とした質問応答システムのプロセスについて整理する．

2.2 コミックを対象とした質問応答技術

1 章でも述べたように，コミックを対象とした質問応答システムでは，テキスト情報ではなくコミックのコマやストーリーの一部分を提示する方が適切な場面が存在する．システムがコマやストーリーの一部分を回答と認識するためには，コミックに含まれる情報 (e.g., キャラクタ情報，セリフ) を構造化し，機械で計算可能にする必要がある．例えば計算が可能になることで，「名探偵コナンで主人公がスケートボードに乗っているシーンが見たい」といった要求に対して，システムは「主人公」と「スケートボード」などの情報がタグ付けされているコマを探し出し提示することが可能になる．コミックの書誌に関する情報を構造化する取り組みは，野村ら [6] や，三原ら [9] によって行われており，Wikipedia[1] や DBpedia[2] を用いることで可能であるとされている．

一方，コミックの内容に関する情報は，上記の情報源からのみでは十分ではないため，コミックの各コマに記載されている各々の情報をより詳細に構造化する必要がある．水戸らは，人手でこれを行うことで，様々な質問に回答するための基盤を構築することを試みている [8]．このように構造化されたデータは質問応答システムに限らずコミックに対する新たなサービスの創出につながると期待されるが，全てのコマに対して情報を人手で付与していくのでは，非常にコストがかかってしまうため，コミックの情報を自動で抽出することが求められる．こうしたコミックの内容情報を自動抽出する方法としては画像認識の利用が見込まれる．谷らは，画像認識技術を用いてコミックの登場キャラクタを認識，識別する手法について検討を行っている [13]．現状のキャラクタ識別の正解率は必ずしも高いとは言えないものの，この技術が発展することでコミック内の特定の情報 (e.g., キャラクタ名，アイテム) を自動で抽出・識別し，アノテーションとしてコマに付与できると期待される．

コミック質問応答は，こうした技術が利用可能であるという前提の下で進める．当面は，水戸らの研究に倣って，構造化されたコンテンツ情報を人手で用意し，それを用いて研究を進めている [8]．

質問応答システムでは，まずユーザが入力した検索クエリが何に関する質問であるのかを判断する．一般的な質問応答のタイプ分類に関する研究は行われているものの (e.g., [14])，1 章でも述べたようにコミックはテキスト情報と画像情報が相補的に利用されているマルチモーダルなコンテンツであるため，一般的な質問応答システムの想定する質問タイプ分類に該当しないような質問タイプが出現する可能性がある．そこで次節では，コミックを対象とした質問のタイプ分類に関する取り組みについて詳述する．

2.3 コミックを対象とした質問タイプ分類

現在，コミックに関する情報は電子書籍販売サイト (e.g., コミックシーモア[3]) などから獲得できる．こうしたサイトでは，一般的に，コミックの表題や著者名，出版社名といった書誌情報による検索が可能である．しかし，コミックの中の特定のシーンを探したい，コミックの内容を手がかりにして表題や著者名を探したいと

[1] http://www.wikipedia.org
[2] http://www.dbpedia.org
[3] http://www.cmoa.jp

いう情報要求には，現状では必ずしも応えられていない [4].

こうした要求は「Yahoo! 知恵袋[4]」や「教えて! goo[5]」などのインターネット上の質問サイトなどで質問することにより，ある程度解決することが可能である．しかし，回答を得るのに時間を要したり，回答が得られない場合も少なくない．このような問題を解決するために，福田らは，2.1 節で述べた質問応答技術の枠組みを採用し，コミックコンテンツに適用するためにユーザから与えられる質問のタイプ分類について検討している [11]．具体的には，「Yahoo! 知恵袋」と「教えて! goo」から各々 30 個，コミックに関する質問を収集し，文中に出現する疑問詞や手がかり語に着目した質問タイプ分類を人手で行っている．以下に質問タイプを示す．

- 位置に関する質問

  特定のシーンや話など，質問対象が収録されている巻数や話数を問う質問

- 登場人物に関する質問

  登場人物の外見や所属など，コミックの設定に関する質問

- ストーリーに関する質問

  作品全体や単行本 1 巻分などに関するストーリーの具体的な内容を問う質問

- 作品のタイトルに関する質問

  いくつかの手がかりとなる項目を挙げて，それらを満たす作品のタイトルを問う質問

- その他に関する質問

  上記のタイプに当てはまらない質問

上記の分類は，コミックに対応した質問タイプ分類ではあるが，人手で分類しているため，この質問タイプ分類が妥当であるのかを評価する必要がある．そこで，次章では，機械学習を用いてコミックの質問タイプを自動で分類することを試みる．

3 実装と事前評価

3.1 学習データの収集

福田らは，質問タイプの傾向を判断するために，「Yahoo! 知恵袋」と「教えて! goo」から計 60 個の質問文を収集し，それに基づいて質問タイプの分類を行なっているが，位置に関する質問が 52 個 (86.7%) と全体に占める割合が多く，質問タイプが偏っている．

そこで本稿では，質問数のバランスをとるために「Yahoo! 知恵袋」と「教えて! goo」から合計 180 個の質問文を収集した[6]．なお，今回収集した質問文は全て正解が一意に決定する Factoid 型質問文を対象とし，正解が一意に決定しない主観による回答が求めれる Non-Factoid 型質問文は収集の対象外とした．加えて，質問文に含まれる，質問の内容とは関係ない文章や単語 (e.g.,“こんにちは”，“回答よろしくお願いします”，“あまりはっきりとは覚えてないんですけど”) は，分類結果への影響を考慮して事前に取り除いた．

3.2 自動分類器を用いた検証

本稿では，質問タイプの自動分類に SVM (Support Vector Machine) を用いた．本稿では，機械学習ライブラリの scikit-learn[7]を用いて分類器を作成し，評価を行った．パラメータはデフォルト (C=1.0) のまま用いた．また，カーネルの種類には，線形カーネル関数を利用した．今回の実験では，収集した質問文を形態素解析器 MeCab[8]により形態素解析して得られた形態素を SVM の素性とした．

3.1 節で収集した質問を用いて福田らの質問タイプ分類の評価を行うために， 5 分割交差検定 (5-fold cross validation) を行った．分類結果の精度 (precision)，再現率 (recall)，F 値を表 1 に示す．表 1 から，位置に関する質問と作品のタイトルに関する質問に対する F 値が高い事が確認された．これは，“何巻ですか？ ” や “タイトルを教えて下さい” などといった特定の表現が多数の学習データに含まれていたからであると推測される．一方，登場人物に関する質問やストーリーに関する質問に対する F 値は低かった．これは，質問の記述形式が各々異なることが原因だと考えている．

今回，“その他” に分類された質問数は全質問文 180 個のうち 43 個であり，高い割合を占めていた．“その他” に分類された質問文の中には，コミックの発売日を問う質問や，福田らが考案した質問タイプ分類には属さないコミックの内容に関する質問 (e.g.,“〜はどういう意味ですか？ ”) などが複数確認された．このことから，福田らが考案した質問タイプでは，想定される質問を分類するには不十分であることが示唆された．そのため次節では，先行研究および今回収集した質問文をもとにして，より多様な質問タイプに適応した分類基準の検討を行う．

[4] http://chiebukuro.yahoo.co.jp (2014 年 5 月 20 日存在確認)
[5] http://oshiete.goo.ne.jp (2014 年 5 月 20 日存在確認)
[6] 2013 年 11 月 1 日時点でアクセス可能な質問文を対象とした．
[7] http://www.scikit-learn.org/stable/
[8] http://www.mecab.sourceforge.net

表 1: 先行研究に基づく質問タイプの分類精度

質問タイプ	質問数	精度	再現率	F 値
位置に関する質問	60	0.76	0.87	0.81
登場人物に関する質問	20	0.47	0.45	0.46
ストーリーに関する質問	26	0.52	0.54	0.53
作品のタイトルに関する質問	31	0.79	0.74	0.77
その他に関する質問	43	0.62	0.53	0.57

3.3 質問タイプ分類の再検討

福田らは，コミックに含まれる要素と収集した質問文に含まれる要素から，書誌情報に関する要素 (e.g., 巻数，作品名) と内容に関する要素 (e.g., キャラクタ，セリフ) の 2 つに大別している．本研究もそれに倣い，コミックの質問タイプ分類の際には，まず書誌情報型の質問タイプと内容情報型の質問タイプに分類して考える．本稿では，福田らが提案した質問タイプ分類と栗山らが提案した情報検索型の質問タイプ (サーチエンジンや図書館のレファレンス・サービスを利用して回答を探すことが可能な質問) ，さらに今回収集した質問のうち，これらに該当しないものを考慮して分類を再度行い，最終的に書誌情報型質問 5 タイプ，内容情報型質問 6 タイプの計 11 タイプに分類することとした．以下にその質問タイプを示す．

- 書誌情報型質問
  - 巻数や話数に関する質問
  - 作品名や各話のタイトルに関する質問
  - 発売日に関する質問
  - 掲載誌に関する質問
  - 作者に関する質問
- 内容情報型質問
  - ストーリーの進展 (結果, 過程) に関する質問
  - ストーリーの定義や解釈に関する質問
  - ストーリーの理由，原因に関する質問
  - キャラクタの設定に関する質問
  - オブジェクト，道具，技能の名称に関する質問
  - セリフに関する質問

次章では，上記の 11 個の質問タイプを用いた質問文の自動分類を行い，その分類精度の評価を行う．

表 2: 書誌情報と内容情報に分類した際の分類精度

大分類	質問数	精度	再現率	F 値
書誌情報型質問	103	0.88	0.86	0.87
内容情報型質問	77	0.84	0.82	0.83

表 3: 書誌情報型質問に関する分類精度

質問タイプ	質問数	精度	再現率	F 値
巻数や話数に関する質問	47	0.67	0.83	0.74
作品名や各話のタイトルに関する質問	40	0.56	0.60	0.58
発売日に関する質問	11	1.00	0.73	0.84
掲載誌に関する質問	3	–	–	–
作者に関する質問	2	–	–	–

4 評価と考察

本稿では，質問タイプ分類を 2 段階で行うこととした．まず，その 1 段階目として 書誌情報型質問と内容情報型質問の 2 つに分類した際の精度を評価した．分類結果を表 2 に示す．結果を見てみると，精度，再現率共に書誌情報型質問の方が高いことが確認された．また，書誌情報型質問，内容情報型質問のどちらとも F 値が高いことが確認された．

次に，書誌情報型質問の 5 分類と内容情報型質問の 6 分類の計 11 分類での分類を行った．ただし，今回の分析を評価するために 5 分割交差検定を用いたため質問数が 5 に満たない質問タイプが混在していると結果に影響を及ぼす可能性がある．そこで本稿では，質問数が 5 以下である掲載誌に関する質問と作者に関する質問を除く 9 種類の質問タイプで評価を行った．結果を表 3 と表 4 に示す．表 3 を見ると発売日に関する質問タイプの精度が良かった．これは，表 1 の結果と同様に，“〜の発売日はいつですか” といった特定の文末表現が多くの質問文に含まれていたのが要因であると考えられる．

一方，表 4 を見るとストーリーの定義に関する質問とオブジェクト，道具，技能の名称に関する質問の F 値が 0 だった．これは，各質問文に特徴的な表現が多く，加えて質問数が少なかったため，上手く分類ができなかったのが原因であると考えられる．これらの質問タイプに関しては，今後該当する質問文を増やし，再度検討を行う必要がある．

また，書誌情報型質問と比較して，内容情報型質問の精度が悪かった．その原因として，内容情報型質問の質問文中に口語表現が多く特徴的な単語の割合が高かったことが考えられる．今後は，精度を改善するために，口語的表現を文語的表現に修正したものを学習データとする手法と固有表現にタグ付けを行い，質問文中に含まれる名詞情報を質問タイプ分類の判断基準に用いる手法を採用していきたいと考えている．

表 4: 内容情報型質問に関する分類精度

質問タイプ	質問数	精度	再現率	F 値
ストーリーの進展 (結果，過程) に関する質問	20	0.36	0.45	0.40
ストーリーの定義に関する質問	7	0.00	0.00	0.00
ストーリーの理由や原因に関する質問	11	0.38	0.27	0.32
キャラクタの設定に関する質問	21	0.52	0.57	0.55
オブジェクト，道具，技能の名称に関する質問	10	0.00	0.00	0.00
セリフに関する質問	8	0.43	0.38	0.40

5 終わりに

本稿では，コミックの質問に対応した質問応答システムを実現するための基礎検討として，コミックに関する質問を収集し，質問タイプ分類を試みた．先行研究の 5 分類で事前評価を行った所，その他の質問に分類される質問が多かったため，先行研究の質問タイプ分類と今回収集した質問文を参考にして新たな質問タイプ分類を行い評価を行った．今後は，話し言葉が多い質問文を書き言葉に変換する等の処理を行い，より分類精度の向上を目指す．また，今回は，Factoid 型質問文を対象に扱ったが，今後は，NonFactoid 型質問文も対象にして分析を行っていきたい．

6 謝辞

本研究は挑戦的萌芽研究 (課題番号:24650040) の助成を受けた．記して謝意を示す．

参考文献

[1] A. Morozumi, S. Nomura, M. Nagamori, and S. Sugimoto. Metadata framework for manga: A multi-paradigm metadata description framework for digital comics. In *Proc. International Conference on Dublin Core and Metadata Applications 2009*, pp. 61–70, 2009.

[2] T. Tanaka, K. Shoji, F. Toyama, and J. Miyamichi. Layout analysis of tree-structured scene frames in comic images. In *Proc. 20th International Joint Conference on Artificial Intelligence*, pp. 2885–2890, 2007.

[3] J. Weizanbaum. A computer program for the study of natural language communication between man and machine. In *Communications of the ACM*, Vol. 9, No.1, pp. 36–45, 1966.

[4] 松下光範. コミック工学の可能性. 第 2 回 ARG WEB インテリジェンスとインタラクション研究会, pp. 63–68, 2013.

[5] 磯崎秀樹, 東中竜一郎, 長田昌明, 加藤恒昭. 質問応答システム. コロナ社, 2009.

[6] 野村聡美, 両角彩子, 永森光晴. マンガのためのメタデータモデルを目指したマンガのアーキテクチャ分析. 第 36 回ディジタル図書館ワークショップ, pp. 3–14, 2009.

[7] 石井大祐, 山崎太一, 渡辺裕. マンガ上のキャラクター識別に関する一検討. 情報処理学会第 75 回全国大会 (分冊 2), pp. 71–72, 2013.

[8] 水戸拓実, 白井涼子, 波多野賢治, 松下光範. コミックデータ内関係抽出のためのデータ・フォーマットの提案. 第 2 回 ARG WEB インテリジェンスとインタラクション研究会, pp. 71–72, 2013.

[9] 三原鉄也, 永森光晴, 杉本重雄. ディジタルマンガにおけるストーリー構造とビジュアル構造を表すメタデータモデル. 情報処理学会研究報告, Vol. 2011-FI-104, No. 9, pp. 1–8, 2011.

[10] 池野篤司. 質問応答システム 情報検索と情報抽出の頂点へ , 技術報告 2. 沖テクニカルレビュー, 2004.

[11] 福田美沙紀, 白水菜々重, 松下光範. コミックを対象とした質問応答技術のための基礎検討. 人工知能学会ことば工学研究会資料, SIG-LSE-C003, pp. 57–62, 2012.

[12] 野中俊一郎, 沢野拓也, 羽田典久. コミックスキャン画像からの自動コマ検出を可能とする画像処理技術「gt-scan」の開発. In *FUJIFILM RESERCH & DEVELOPMENT*, No. 57, pp. 46–49, 2012.

[13] 谷悠, 白水菜々重, 松下光範. コミックコンテンツにおける登場キャラクター抽出のための基礎検討. 情報処理学会第 75 回全国大会 (分冊 4), pp. 889–890, 2013.

[14] 栗山和子, 神門典子. Q&A サイトにおける質問と回答の分析. 情報処理学会研究報告, Vol. 2009-DBS-148, No. 19, pp. 1–8, 2009.