T-Scroll：時間的トピックの推移をとらえる可視化システム

T-Scroll: A Visualization System for Temporally Changing Topics

長谷川 幹根♡ 石川 佳治♢

Mikine HASEGAWA Yoshiharu ISHIKAWA

インターネット上では，ニュースなどの大量のテキストデータの配信が日々なされている．本論文では，このようなテキストデータにおける，時間的なトピックの推移をとらえるための可視化システム T-Scroll について述べる．本システムは，下位の時系列的な文書クラスタリングシステムのクラスタリング結果をもとに，クラスタの関連を巻き物（scroll）状に提示する．本論文では，システムのアイデア，機能，実現手法等について述べる．

On the Internet, delivery of a large amount of documents such as news articles is continually performed everyday. In this paper, we describe an information visualization system T-Scroll to show the transition of topics contaned in such documents to the user and to provide an overview of their trends. The system is built on a clustering system for time-sereis of documents and presents relationships between clusters like a scroll. This paper describes the idea, the functions, and the implementation of the system.

1. はじめに

インターネット上の情報提供・配信サービスの進展により，今日では，ネットワークを介したニュース配信が盛んに行われている．それに伴い，大量の情報を要約しフィルタリングするための，オンラインテキスト情報処理の重要性がさらに増してきており，時々刻々と配信される時系列的な文書データに適した情報の要約と提示に関する新たな技術の開発が求められている [1].

このような背景を受け，本研究では，一般のユーザが大量のニュースのトピックの大まかな推移を容易に把握できるようにするためのユーザインタフェースである **T-Scroll**（Topic/Trend-Scroll）システムの開発を行っている．T-Scroll は文書クラスタリングシステムの上位に位置し，その出力を利用して，クラスタリングされた結果を可視化してユーザに提示する．その特徴は，各時点で得られたクラスタをラベルを付与して時間軸上に配置し，クラスタ間の関連性を表すリンクを示すことで，トピックの流れを表す点にある．画面上にクラスタリングの結果を巻き物上に表示することから，システムを T-Scroll と呼んでいる．あるトピックに興味をもったユーザは，対話的な操作により，必要に応じてより詳細な情報を得ることが可能となる．

2. 新規性に基づく時系列文書のクラスタリング

本研究が基礎とするのは，[4, 5] において提案されている，新規性に基づく文書クラスタリング手法である．その特徴は以下の 3 点である．

1. 類似度計算において，文書の内容の類似度だけでなく文書の新規性も考慮することで，新規性の高い文書により着目したクラスタリング結果を導出する．ポイントとなるのは，文献書誌学で用いられる**老化**（aging）の概念を取り込んだことにあり，若い文書（入手されて間もない文書）ほど，クラスタリング結果に与える影響が高くなるようにしている．
2. 新たに文書の追加の際にはインクリメンタルな更新処理を行い，更新コストを削減している．クラスタリングのアルゴリズム自体は ***k*-means 法**に基づき，それを拡張することでインクリメンタルな処理を実現している．
3. 上述のように本手法では老化の概念を導入しており，文書が古くなると，他のどの文書とも類似しなくなり，外れ値（outlier）となる．外れ値がある場合にクラスタリング結果を悪化させないための処理が工夫されている．また，十分古くなった文書は，寿命に達したとされ，自動的にクラスタリングの対象から削除される．

[4, 5] で用いられた影響力の逓減モデルでは，文書の価値（重み）が時間の経過にしたがって指数的に逓減していくと想定し，文書 d_i に対する文書の重みを $dw_i = \lambda^{\tau - T_i}$ $(0 < \lambda < 1)$ と与える．ただし，τ は現在の時刻を表し，T_i は文書 d_i が入手された時刻を表す．λ は文書の影響力の逓減の度合いを表すパラメタである．一方，n 個の文書からなる文書集合 $d_1, \ldots, d_n$ の文書の重みの総和を $tdw = \sum_{l=1}^{n} dw_l$ で与え，文書 d_i の文書集合中での生起確率を $\Pr(d_i) = dw_i / tdw$ という主観確率で定める．この確率は，古い文書ほど値が小さくなり，古い文書を考慮の対象から外す（忘却する）というアイデアを表現している．

文書の類似度は，上記の式や他の仮定をもとに確率的なモデリングに基づいて導出される [4, 5]．その一般形は

$$sim(d_i, d_j) = \Pr(d_i)\Pr(d_j)\frac{\boldsymbol{d}_i \cdot \boldsymbol{d}_j}{len_i \times len_j} \quad (1)$$

であり，文書ベクトルの内積を文書長の積で割ったものに各文書の生起確率を掛けたものとなる．よって，この文書類似度は，単に文書どうしが類似しているかどうかだけでなく，各文書がどの程度古いかも考慮し，十分古くなった文書は他のどの文書にも類似しなくなるという性質を有している．このような類似度をクラスタリングに用いることにより，文書の新規性を重視したクラスタリングの実現を図っている．

3. T-Scroll システムの概要

3.1 システムの特徴

本研究で開発を進めている **T-Scroll** (Topic/Trend-Scroll) システムの特徴は，主として以下のようになる．

1. 継続的なクラスタリングにより得られた各時点のクラスタリング結果を時間軸上にトピックを表すラベルとともに表示し，各時点における主要なトピックを把握可能とする．ニュース記事などのトピックやトレンドの流れが巻き物のように表示されることから，本システムを T-Scroll と呼んでいる．
2. 興味のあるクラスタを選択することで，より詳細な情報（関連キーワードのリスト）や元記事を対話的に参照することが可能である．
3. ある時点で得られたクラスタ集合に対し，一つ前の時点で得られたクラスタ集合から，関連度の強さに応じてリンクを張ることで，隣接する時刻におけるクラスタ間の関連の把握を容易にする．
4. ユーザインタフェース上に表示する時間軸の刻み幅をユーザの指定により調整可能とすることで，要求に合わせた詳細度で分析が行える．特に，時間軸の刻み幅を広くとり，ト

♡ 学生会員 名古屋大学工学部電気電子・情報工学科.
hasegawa@db.itc.nagoya-u.ac.jp
♢ 正会員 名古屋大学情報連携基盤センター
ishikawa@itc.nagoya-u.ac.jp

レンドを大まかにとらえる粗視化の機能が重要であり，これは，OLAP (On-Line Analytical Processing) におけるロールアップ（roll-up）の機能に対応づけることができる．

3.2 システムの概要

図 1 に，T-Scroll システムのインタフェースの概念図を示す．図は，10 月 1 日から 1 週間刻みで 10 月 15 日までのクラスタの流れを表示している様子を示している．インタフェース上では左から右に時間が流れており，画面下部のスライドバーにより，前後の時点に移動することも可能である．画面上で同じ縦の点線上にある楕円は同じ時点で得られたクラスタの集合を表している．

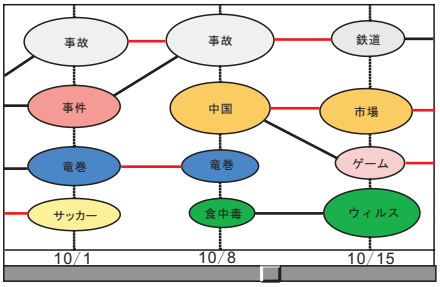

図 1: T-Scroll システムの概念
Fig. 1: Concept of T-Scroll System

クラスタ上のラベルは，クラスタ中の文書に含まれる語で，スコアが最大のものを選択して表示する．いくつかのスコア付けを比較した結果，現在の実装では，クラスタ C_p における語 t_j のスコアを $score(t_j) = \sum_{d_i \in C_p} \Pr(d_i) tf_{ij}$ で求めている．つまり，クラスタ内の各文書について，語 t_j についての語頻度（term frequency）tf_{ij} を，その文書の重み $\Pr(d_i)$ と掛け合わせ，その総和をとっている．なお，クラスタ上に複数の単語（たとえばスコアが上位 3 件の語）を並べて提示することも考えられるが，実システムで検討したところ，画面表示が煩雑になるため 1 語だけを選んでいる．

楕円の面積はクラスタに含まれる文書の数の量に対応しており，トピックの規模を示している．図で示されるように，一部のクラスタ間には左から右にリンクが張られている．これはクラスタ間の関連性の深さを示している．クラスタ間の関連度は $csim(C_i, C_j) = |C_i \cap C_j|/|C_i|$ という式により定義する．クラスタ C_i に含まれる文書がクラスタ C_j にどれだけ含まれているかを調べることにより関連性の深さを測っている．1 つのクラスタから 0 個以上のリンクが出ることを許し，トピックの消滅（0 個のリンクで表現）や分岐（複数個のリンクで表現）を表す．

4. 実装システムの機能

4.1 インタフェース画面

図 2 では，2006 年 10 月 1 日から 1 週間刻みで 12 月 31 日までの時間的トピックの推移を表示した例を示している．楕円はそれぞれのクラスタを表しており，それぞれ 20 個ずつにクラスタリングされている．前節で述べたように，楕円の大きさはクラスタのサイズを大まかに反映する．

クラスタのサイズだけでなく，クラスタの質の良さも把握できるようにするため，T-Scroll ではクラスタの質の高さを色分けして表示する．具体的には，楕円の輪郭の線の色により，クラスタの質の良さを表現する．可視光線のスペクトル分解を参考にし，赤に近いほどクラスタの質が高く，紫に近いほどクラスタの質が低いことを意味する．クラスタ C について，その品質のスコアを，

$$quality(C) = |C| \cdot avg\_sim(C) \quad (2)$$

$$avg\_sim(C) = \frac{1}{|C|(|C|-1)} \sum_{d_i, d_j \in C, d_i \neq d_j} sim(d_i, d_j) \quad (3)$$

と与える [5]．ここで $|C|$ はクラスタ C 中の文書数を表し，

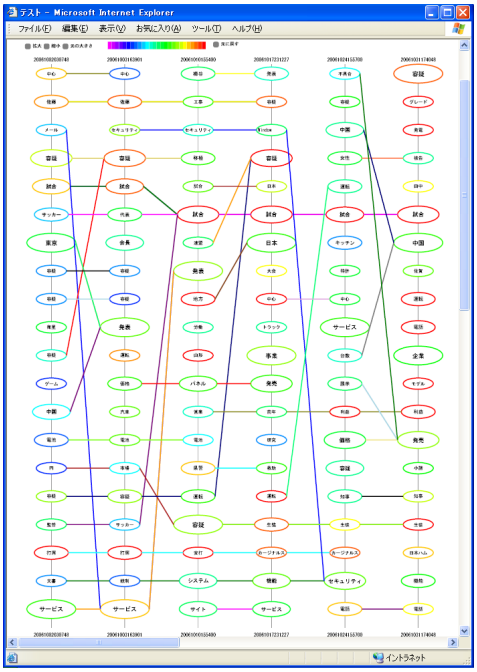

図 2: T-Scroll のインタフェース画面
Fig. 2: Screen Shot of T-Scroll (1 week basis)

$avg\_sim(C)$ はクラスタ内の文書の平均類似度を表している．すなわち，$quality(C)$ は，文書数が多いだけでなく，クラスタ内の文書が互いに似ている場合に大きい値をとるようなスコアとなっている．[5] のクラスタリング処理では，クラスタリングの結果生じるクラスタ集合において，それらの品質の総和が最大となることを目標としてクラスタリングを行う．

3 節で述べたように，クラスタ間のリンクは，クラスタ間の関連度が大きいことを表し，ある閾値以上の関連度についてリンクを作成している．

4.2 クラスタの詳細情報

図 2 のように，クラスタに対するラベルとして 1 つのキーワードを与えるだけでは，クラスタ内容を判断するのが困難な場合もある．そこで本システムでは，クラスタの内容を容易にブラウズできる機能も提供している．クラスタ上（楕円上）にマウスカーソルが乗ると，そのクラスタに関連の深い複数のキーワードが表示される．実行した様子を図 3 に示す．クラスタ内の単語のうち，スコアが上位 20 位のものを順に表示している．

図 3: クラスタのキーワードリストの表示
Fig. 3: Keyword List Display for a Cluster

上記のようなキーワード表示機能によってクラスタの内容はわかるが，実際にクラスタに含まれる文書はわからない．よって，本システムでは更に，クラスタの上をクリックすることでクラスタに含まれる文書を表示する機能も実現している．実行の様子を図 4 に示す．図 4 では，クラスタに含まれる文書のうち発行日時が

新しいもの上位 10 位のタイトルを表示している．文書の内容はタイトルをクリックすることによって表示される．また，詳細情報をクリックすることにより，クラスタに含まれるすべての文書を表示する機能も実装している．

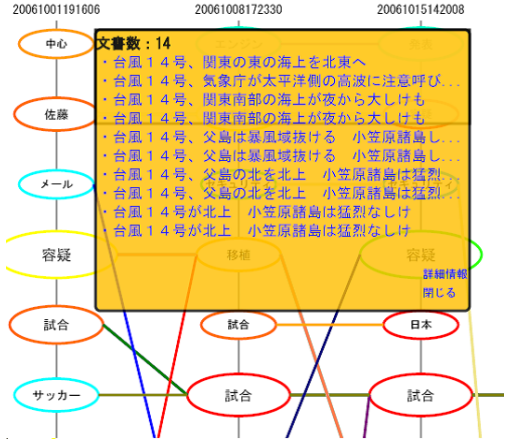

図 4: クラスタ内の文書の表示
Fig. 4: Document Display within a Cluster

5. システムの実装

本システムは以下の図 5 のような構成をしている．本システムは，新規性に基づく時系列文書のクラスタリングのプログラム [5] と連携し，その出力を利用する形で構築している．各時点で取得された新たな文書集合をバッチ的に与えることで，その時点の最新のクラスタリング結果を出力する．

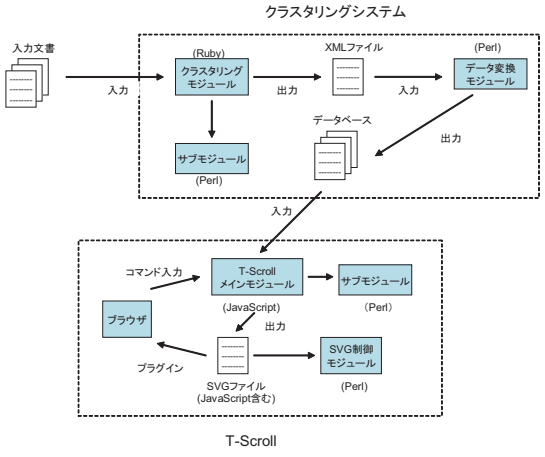

図 5: システム構成図
Fig. 5: System Organization

本実験において対象とした情報源は，RSS データを提供しているニュースサイトである nikkeibp.net, asahi.com, sportsnavi.com（サッカー・野球）の 4 つのサイトである．情報収集は 2 時間おきに行っている．それぞれの RSS サイトにアクセスし，前回情報収集した時から更新された情報について，リンク先などの必要な情報を取得する．次いで，取得したリンク先情報をもとに，サイトにアクセスしウェブページから記事の本文を抽出する．

T-Scroll のメインモジュールは JavaScript で記述されており，Web ブラウザ内に読み込まれ動作する．ユーザインターフェースに関する一部の処理は JavaScript および AJAX の機能を用いて実現している．ユーザから対象の期間や分析の時間間隔の入力を受けた後でインタフェース画面を表示するが，そのためには，メインモジュールから Perl で作成されたサブモジュールを呼び出すことになる．実際にはこのサブモジュールがクラスタリング結果の XML ファイルを読み込み，ユーザの指定に応じて内容を解析し，インタフェース画面に表示するための SVG 形式のファイルを作成する．作成された SVG ファイルはブラウザに即座に読み込まれ，図 2 に示したインタフェース画面が表示される．SVG ファイル中には JavaScript のコードが埋め込まれており，その中から必要に応じて Perl により記述されたモジュールが実行される．

6. システムの評価

6.1 システム利用による評価

まず，実際にシステムを利用した筆者により得られた知見を報告する．今回は，5 節で述べた 4 つのサイトからの記事を対象としており，1 日あたり平均しておよそ 100 件のニュース記事が取得されている．設定により，各時点において 20 件のクラスタが作成され表示されている．表示の対象とする期間については，長期（例：3ヶ月以上）に設定することはあまり有効とはいえなかった．トピックの推移は 1〜2ヶ月程度ぐらいの範囲でとらえる方が分かりやすいという点と，長期の場合には表示が煩雑になり，また，インタフェースの動作が重くなるためである．

時間間隔の設定については，1 日刻みで表示した場合には比較的単調な表示となる．その様子を図 6 に示す．この図は，12 月 1 日から 1 日刻みで 12 月 10 日までトピックの推移を表示している．利用した印象としては表示が冗長であるという感触を得た．これは，1 日程度では大きなトピックの変化がないためである．一方，1 週間刻みで表示した場合（図 2 参照）には，トレンドを把握するという意味ではより適切な表示であると感じられた．インタフェースの表示においても，クラスタ間のリンクの交差などが見られ，視覚的には面白いものとなっている．ただし，たまにリンクが張られている隣接するクラスタでトピックがずれていること，すなわちトピックドリフトが見られた．

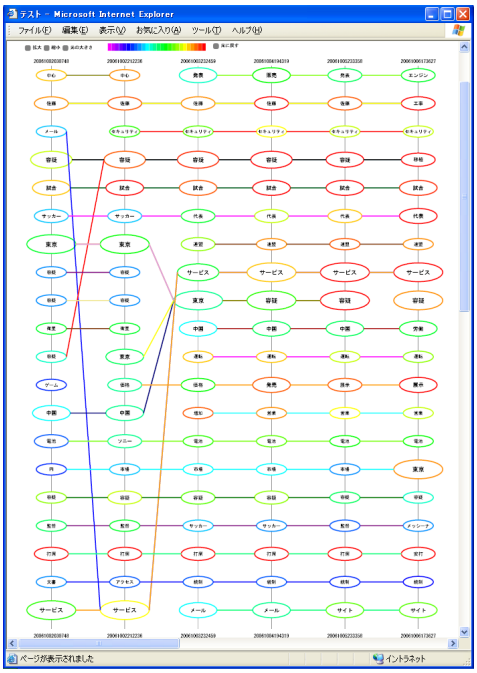

図 6: T-Scroll 全体図（1 日刻み）
Fig. 6: Screen Shot of T-Scroll (1 day basis)

2006 年度後半の実際のデータについて観測できたさまざまな知見については，紙面の都合によりここでは省略する．詳細は [2] を参照いただきたい．

6.2 クラスタのトレンド評価

本節では，2006 年後半に実際に起きた出来事の流れとクラスタのトレンドを比較し，T-Scroll のクラスタのトレンドの正確性を評

価する．評価にあたり，各クラスタの内容判断は，クラスタに含まれる文書のうち発行日時が新しいもの上位 10 件までを対象とし，上位 10 件までに対象とする出来事に対する記事がどれくらいの割合で含まれているかを評価の値（**トレンド値**と呼ぶ）として用い，クラスタのトレンドとしてグラフに表し評価する．

2006 年 10 月 1 日から 12 月 31 日までの様々な出来事に対するクラスタのトレンドの評価を行ったが，ここでは例として「知事談合」に関するクラスタのトレンドの評価を示す．知事談合に関する主要な出来事は，以下のようになっている．

9 月 27 日頃：福島県知事談合問題発生
10 月 7 日頃：和歌山県知事談合問題発生
10 月 23 日：福島県知事逮捕
11 月 15 日：和歌山県知事逮捕
11 月 16 日頃：宮崎県知事談合事件発生
12 月 8 日：宮崎県知事逮捕

図 7 に知事談合に関するクラスタのトレンドを示す．10 月 1 日から 12 月 30 日まで 3 日ごとのクラスタのトレンドを示している．実際の事件のトピックのトレンドときわめて整合した結果となっている．詳細な分析は [2] で述べている．

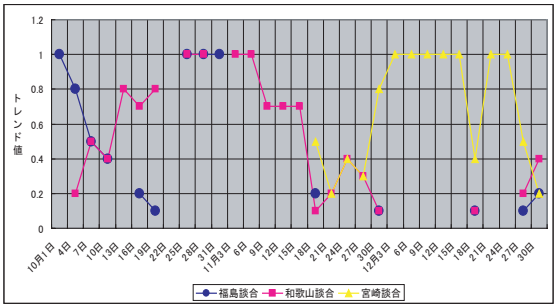

図 7: 「知事談合」に関するトレンド
Fig. 7: Trend Plot for “Chiji-Dangou”

3ヶ月のクラスタのトレンドの評価より得た知見を以下に示す．

- スポーツや自然災害などのクラスタのトレンドは時期が去っても高いトレンド値を維持することが多い．これは，スポーツや自然災害などが他の種の記事とあまり類似度が高くないため，ある文書の重みが小さくなっても程度の記事が消滅するまで残ってしまうと考えられる．
- 政治に関するクラスタのトレンドがほとんど現れない．これは，先に述べた通り今回利用したニュースサイトが政治に関する記事が少なかったことと政治に関する記事は 1 つのクラスタに集まりやすいためだと推測される．
- 裁判の判決など事前に起こる時期が分かっている出来事は，発生よりも前から低いトレンド値でクラスタのトレンドが現れることが多い．また，地震や事件など先に予測できない出来事は，急にクラスタのトレンドが現れることが多い．
- 多くのクラスタのトレンドが事件などの発生時期よりも遅れる．これは，まとまったクラスタとして現れるためには，それなりの記事の量が必要であるためであると考えれる．

T-Scroll のクラスタのトレンドは，事件などの発生や時期が過ぎた後に正確でないトレンド値を記録することがあるが，最もホットな時期にはクラスタのトレンドの中で最高値を記録することがほとんどである．これにより，T-Scroll は大まかなトピックのトレンドをとらえるのには有効であると評価できる．

7. 関連研究

ThemeRiver [3] は，トピックの流れを川に見立てて表示を行う可視化システムであり，川が画面の左から右に流れるような表示を用いる．川の中にいくつかの色分けされた流れが表示されており，これが一つ一つのトピック（テーマ）に対応している．また，川の幅は各時点における記事の量を表している．トピックの流れを左右にスクロールするインタフェースで表現するという点では T-Scroll と共通しているが，クラスタリングを用いているわけではない．視覚的なインパクトはあるが，トピックの推移は表現できず，複数の時間間隔での表示なども可能でない．大まかなトレンドの把握には利用可能であるが，実際に時系列的な文書データを分析的にブラウズするには，必ずしも強力なツールではない．

Swan と Allan は，トピックを表現する timeline を表示するインタフェースを提案した [6]．指定された期間における時系列的な文書を分析して，継続して出現するトピックを検出し，画面上に時区間を表す棒状の表示（timeline）を提示する．また，timeline には併せてキーワードが表示される．検出されたトピックごとに timeline が提示されるため，ユーザは画面を眺めることでトピックがどの期間に見られるかを把握できる．クラスタリングではなく，統計的指標を用いてトピックの検出を行っており，主要なトピックとその期間を提示することに焦点を当てている．その点に関しては T-Scroll より優れている面もあるが，トピック間の関連や，複数の時間間隔による分析機能はない．

8. まとめと今後の課題

本論文では，時系列的な大量のオンライン文書のトピックの変遷・推移を対話的に分析するためのインタフェースである T-Scroll システムの特徴，機能，構成，そしてその評価について述べた．今後の課題としては，日本語以外の記事への対応，および，マルチユーザ環境への対応が考えられる．

[謝辞]

本研究の一部は，文部科学省科学研究費（19024037），日本学術振興会科学研究費（19300027），放送文化基金，および柏森情報科学振興財団の助成による．

[文献]

[1] J. Allan, editor. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer, 2002.

[2] 長谷川幹根, 石川佳治. T-Scroll：時間的トピックの推移をとらえる可視化システム. 電子情報通信学会データ工学ワークショップ（DEWS2007）, 2007.

[3] S. Havre, et al. ThemeRiver: Visualizing thematic challenges in large document collections. *IEEE Trans. on Visualization and Computer Graphics*, Vol. 8, No. 1, pp. 9–20, 2002.

[4] Y. Ishikawa, Y. Chen, and H. Kitagawa. An on-line document clustering method based on forgetting factors. In *Proc. ECDL*, pp. 332–339, 2001.

[5] S. Khy, Y. Ishikawa, and H. Kitagawa. A novelty-based clustering method for on-line documents. *World Wide Web Journal*, 2007. (to appear).

[6] R. Swan and J. Allan. Automatic genration of overview timelines. In *Proc. ACM SIGIR*, pp. 49–56, 2000.

長谷川 幹根　Mikine HASEGAWA
2007 年名古屋大学工学部電気電子・情報工学科情報工学コース卒．情報検索の研究・開発に従事．現在，日本製粉 (株) に勤務．

石川 佳治　Yoshiharu ISHIKAWA
名古屋大学情報連携基盤センター教授．データベース，データ工学，情報検索等に興味を持つ．日本データベース学会，情報処理学会，電子情報通信学会，人工知能学会，ACM，IEEE CS 各会員．