3K3-9

トピック連結に基づく文脈再構成のためのトピック遷移分析

Topic Transition Analysis for Context Reconstruction Based on Topic Concatenation

加藤 義清[*1] 赤石 美奈[*2] 堀 浩一[*2]
Yoshikiyo Kato Mina Akaishi Koichi Hori

[*1]情報通信研究機構
National Institute of Information and Communications Technology

[*2]東京大学先端科学技術研究センター
Research Center for Advanced Science and Technology, the University of Tokyo

In order to support users in comprehension and utilization of vast amount of information, it is crucial that they should be able to untangle the intertwined context embedded in the information, and to flexibly recontruct information according to their own context. This study aims at realizing a framework that enables reconstruction of information according to context provided by users. In particular, we target at analysis of temporal structure of topic underlying in documents with time attribute. In this paper, we first introduce a framework for reconstructing context in information based on *topic concatenation*. Then, we propose a method for dynamic analysis of topic structure in documents with time attribute.

1. はじめに

膨大な情報を適切に理解し，活用するためには，同じ情報を様々な観点からとらえ，複雑に絡み合った文脈を解きほぐし，ユーザの要求に応じて柔軟に情報を再構成していくことが必要である．本研究では，時間属性付き文書集合に内在するトピックの時間構造を様々な観点から分析し，更にはユーザ自身の与える文脈に沿った形での情報の再構成が可能となる枠組みの実現を目的とする．本稿では，トピック連結に基づく文脈再構成の考え方を提案し，更にそのために必要となる時間属性付き文書集合を対象としたトピック構造の時間変化を分析する手法について報告する．

2. トピック連結に基づく文脈再構成

知識管理では，組織の中で蓄積された膨大な情報をいかにして知識化し活用していくかが問題となっている．そのためには，情報を静的なものとして扱う従来の情報検索やデータマイニングの手法だけでなく，ユーザの状況や要求に応じて，情報を多角的な観点から捉え，再構成し，ユーザの知識創造を支援することが求められている [赤石 06].

本研究では知識創造を支援する具体的な問題として，文書集合が与えられたときに，いかにユーザがその要求に応じて既存の情報を再構成し，問題解決に至るような知識を発見あるいは想像することができるかについて考える．特に，情報を再構成する上で，トピックが重要な基礎単位であり，その連結により新たな知見が生み出されるとの仮説のもとに，トピック連結に基づく文脈再構成という考え方を提案する．以下，具体例を用いながらトピック連結による文脈再構成の意義を示す．

図 1 は，時間属性付き文書集合として，ある人工衛星開発プロジェクトの議事録を対象にして，特定のキーワードで検索された文書を，文書の時間属性順に文書をノードとする連鎖として可視化したものである．この図では 3 つのキーワード，すなわち「電池 問題」「電池 起動」「昇圧」に対応する 3 つの連鎖が表示されている．まずユーザは議事録から人工衛星に搭載される電池の問題に関係した文書を探すために「電池 問題」というキーワードで検索をおこなう．その結果，電池の起動に関する問題について書かれた文書を発見する．そこで,「電池 起動」という新たなキーワードで検索をおこなう．この時,「電池問題」に対応する連鎖は表示されたままで，両方のキーワードを共有する文書に対応するノードをそれぞれの連鎖が共有する形で表示される．この文書を連鎖の連結点と呼ぶことにする．この連結点に着目すると「昇圧」というキーワードが浮かび上がる．そこで更に「昇圧」というキーワードで連鎖を追加すると，その連鎖に含まれる文章に昇圧に関係して起きる電池の起動の問題の現象とその原因についての記述を見つけることができた．

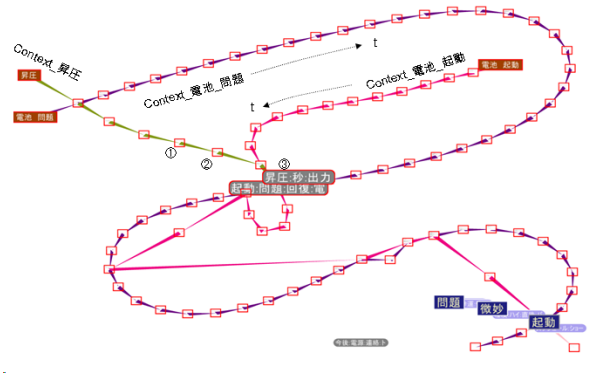

図 1: 人工衛星開発プロジェクトの議事録における複数文脈の相互作用の可視化.

このように，文書集合の中から特定のトピックに沿った文脈を文書の連鎖として取り出していき，更にユーザの気づきによって得られた新しいキーワードから新たな文脈を重ね合わせていくというインタラクションを通して，ユーザは文書集合の

連絡先: 加藤 義清, 情報通信研究機構 知識創成コミュニケーション研究センター 知識処理グループ, 〒 631-0289 京都府相楽郡精華町光台 3-5, Tel:0774-98-6874, Fax:0774-98-6960, ykato(at)nict.go.jp

中から有用な知識を発見することが可能となる．

トピック連結に基づく文脈再構成においては，2 つのことが重要となってくる．1 つ目は，文脈の基本単位である．上述の例ではキーワード，文書，および文書の順序関係に基づいた連鎖を一つの単位として連結していった．しかし，これはあくまでも 1 つの例であって，他にも基本単位は考えられる．2 つ目は，文脈再構成におけるユーザの気づきである．この例では，ユーザ自身が有用な知識発見につながるキーワードを見つけて文脈の再構成に成功している．いかにユーザの気づきを促すような，ユーザの文脈に関連性の高いキューを提示することができるかが問題となる．

3. トピック遷移解析

トピック遷移解析とは，与えられた文書集合についてそこに現れるトピックが時間的にどのように変化しているかを捉えるための解析である．前節でも述べたように，トピック連結に基づく文脈再構成においては，文書集合中のトピック構造を明らかにした上で，トピック構造の構成要素を用いてユーザの文脈に応じて情報を再構成する．前節で示した例では単一のキーワードをトピックと見なしたが，単一のキーワードの場合，文脈によって意味が異なったりして，必ずしも有効な基本単位であるとは限らない．そこで，文書集合に潜在トピック分析を適用することに得られる潜在トピックを基本単位として利用することを考える．

1. 文書集合に対して潜在トピック分析を適用し，トピックパラメータを推定する．
2. 得られたトピックパラメータに基づいて，各文書のトピック分布を求める
3. 文書を時区間に分けて，同じ時区間に含まれる文書のトピック分布の和から時区間毎のトピック分布を求める．

以上の方法により，各潜在トピックについて，時区間毎のトピックの強度が得られる．

3.1 潜在トピック分析

文書集合に潜在するトピックを分析する方法として，Latent Dirichlet Allocation (LDA) [Blei 03] を用いた．

LDA は多重トピックのテキストモデルである．各文書についてディリクレ分布に基づいてトピックの混合比率 θ が与えられ，更に θ に基づいてトピック z_i の選択，トピック z_i に応じた単語生成確率モデル $p(w|z_i) = \beta_i$ に基づいてテキストが生成されるというものである．変分ベイズ法により近似的にパラメータ推定がなされる．

3.2 時区間毎のトピック分布

LDA により，各文書について潜在トピックの混合比率 θ が求まる．各文書をその時間属性に基づいて時区間毎に分け，各時区間に含まれる文書のトピック混合比率の和を取ることで，各時区間におけるトピック分布を求めることができる．

4. 実験

提案手法を用いて，Web ページを対象にトピック遷移解析を適用する実験をおこなった．データは「テロ特措法」をクエリとして，Google API の期間指定検索機能を利用して，2001 年 9 月から 2007 年 9 月を 1ヶ月毎に期間を区切って検索された約 10,000 ページである．複数の区間に重複して現れるページについては最新の区間にあるもののみを残した．このデータについて LDA 法 [Blei 03] を適用し，潜在トピック分析を施し，得られた潜在トピック毎の強度の遷移を図 2 に示す．各潜在トピックのキーワードは潜在トピックパラメータであるトピック毎の語の出現確率 $p(w|z)$ に基づいて，各トピックに特徴的なキーワードを抽出したものである．

図中の丸印で示したトピックは，クエリの「テロ特措法」と関連が深いと思われるものである.「テロ，措，給油，延長」について，テロ特措法の延長が争点となっていた参議院選挙の時期にトピック強度が強く現れている．また，安倍首相が辞任した後の後継総裁選びの時期に「福田，麻生，選，安部，自民党」というキーワードで特徴付けられる潜在トピックが現れている．本手法による文書集合の俯瞰が有効であることを示唆している．一方で，ノイズと思われるトピックも多く出てきている．Web ページの主要部ではなく，Web ページのテンプレート部分で現れる語が多く含まれるトピックもあり，Web ページの主要部抽出の技術などと組み合わせることを検討する必要がある．

5. 関連研究

LifeLines [Plaisant 96] は，医療記録や少年の非行歴など，個人的な履歴情報を時間軸上に年表形式で表示し，関連情報へのアクセスを提供する．治療行為や，非行歴など現実世界での具体的な事象が表示の対象となっており，本研究の対象とは異なる．

ThemeRiver [Havre 02] は，文書集合に含まれるトピックの時間的変化を川のメタファーにより時間軸上に可視化する手法である．文書集合のトピックを対象とする点は本研究と同じであるが，ThemeRiver は文書集合の全体的なトピックの分布の変化を可視化する手法であり，本研究が対象とする，個別的な文脈の俯瞰，及びそれらの間の関係を捉えることは出来ない．

Thread Arcs [Kerr 03] は電子メールの到着時間および返信関係を用いて，電子メール間の関係を可視化する手法である．本稿では電子メールを例に取り上げたものの，本提案手法では一般的な文書を対象としており，特に返信関係は用いていない．ただし，電子メールによるやり取りを理解するためには，返信関係は重要な要素であると Kerr も述べており [Kerr 03]，電子メール，電子掲示板，ニュースグループ等，返信関係を有する文書を対象とする場合に，それをどのように活用するかについては考慮の余地がある．

6. おわりに

本稿では，知識創造を支援するためのトピック連結に基づく文脈再構成の考え方を示し，そこで必要となるトピック遷移分析の方法を提案した．実験の結果，文書集合の内容を良く表す関連性の高い潜在トピックが得られ，有効性を示唆する結果が得られたと考える．但し，ノイズとなる潜在トピックも混在することが明らかとなった．今後，ユーザにより関連性の高いキューを提示するために，表示する潜在トピックの洗練やキーワードの選択が課題となる．

参考文献

[Blei 03] Blei, D., Ng, A., and Jordan, M.: Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

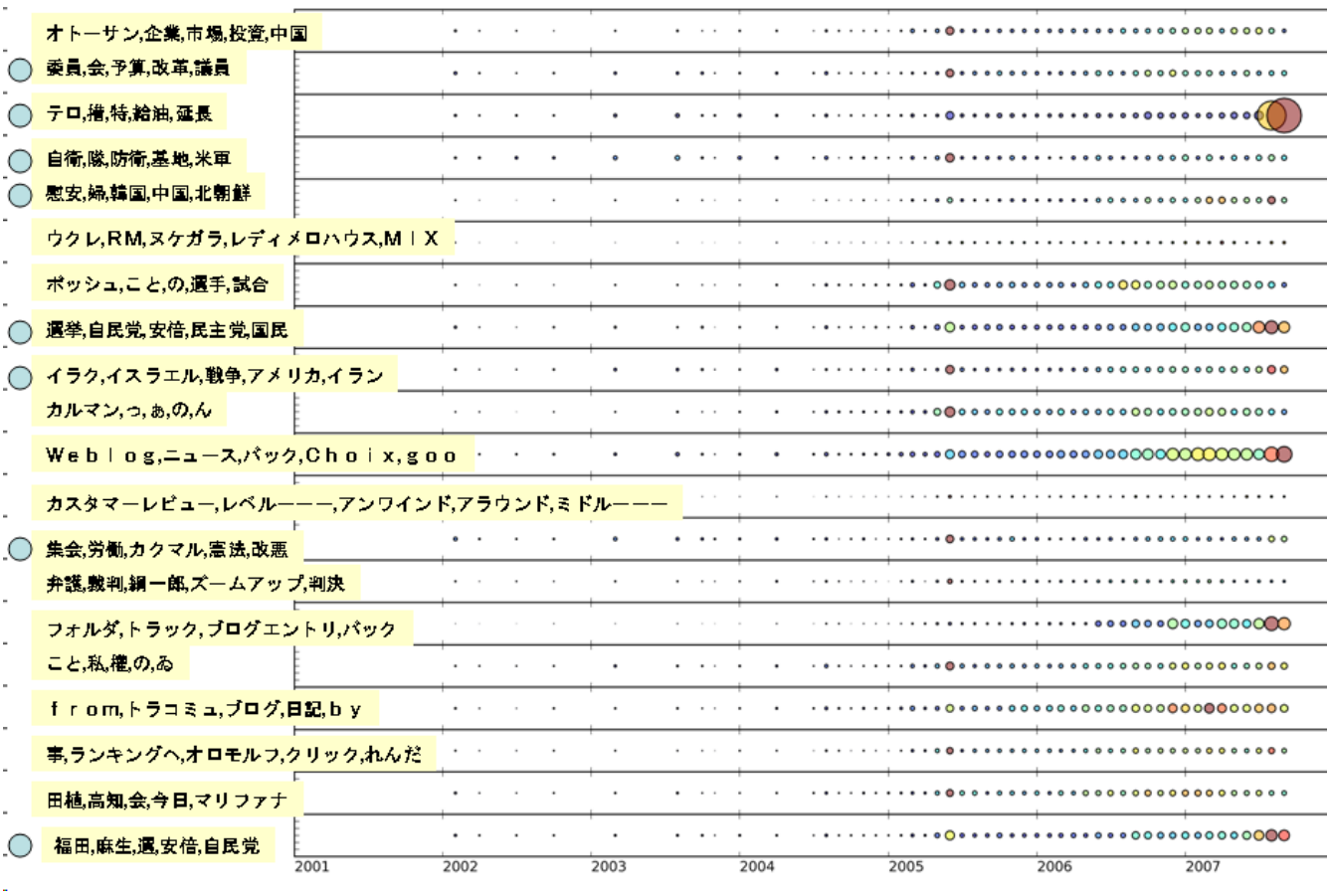

図 2: 「テロ特措法」で検索された 2002 年 2 月から 2007 年 9 月に公開された約 10,000 の Web ページに対するトピック遷移解析の結果.

[Havre 02] Havre, S., Hetzler, E., Whitney, P., and Nowell, L.: ThemeRiver: visualizing thematic changes in large document collections, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 8, No. 1, pp. 9–20 (2002)

[Kerr 03] Kerr, B.: THREAD ARCS: An Email Thread Visualization, in *IEEE Symposium on Information Visualization, 2003 (INFOVIS 2003)*, pp. 211– 218 (2003)

[Plaisant 96] Plaisant, C., Milash, B., Rose, A., Widoff, S., and Shneiderman, B.: LifeLines: visualizing personal histories, in *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, p. 221 ff. (1996)

[赤石 06] 赤石 美奈：文書群に対する物語構造の動的分解・再構成フレームワーク, 人工知能学会論文誌, Vol. 21, No. 5, pp. 428–438 (2006)