3H8-4

文書集合に潜在する多重文脈の相互作用の可視化

Visualizing Interaction of Latent Contexts in Documents

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 this paper, we propose a method for extracting and visualizing latent contexts from documents with time attribute. We describe the technique, present an example in which the technique was applied to a set of e-mails discussing the design of an artificial satellite, and discuss the issues of the technique of current status.

1. 序論

本研究では，文書集合の俯瞰的な理解，および情報アクセス手段のために，文書集合に潜在する複数の文脈の抽出，およびそれらの文脈の間の相互作用の可視化に基づく，文脈に沿った情報アクセスが可能なインタラクション環境の実現を目的とする．

現在，情報通信技術の発展に伴い，アクセス可能な電子文書の数が爆発的に増大している．情報検索技術の研究開発は進んでいるものの，情報検索は利用者の情報利用の要求の一部しか満たすことが出来ない．例えば，良く知られた単一のトピックについて検索エンジンは非常に有効であるが，2 つ以上のトピックについての関連を調べようと思うと，検索と情報閲覧を何度も繰り返しながら，利用者自らの手で関係をまとめる必要があり，不可能でないにしても非常に困難である．

文書に記述された情報を理解する上で，文書が執筆された，あるいは発行された日時などの時間属性は非常に重要な文脈を与える．会議の議事録や電子メールを後から見返す際に，日付を頼りに文書を探したり，日付から文脈を想起して文書の記載内容を理解したりすると言ったことは日頃からおこなわれていることである．更に，近年ブログなど明示的に日付入りの文書が流通し，いつ発行されたかという情報が情報をアクセスする上で非常に重要な情報となっている．

そこで，本研究では時間属性付きの文書集合に着目し，そこから複数の文脈を抽出，可視化することによって，文書集合の俯瞰的な理解と文脈間の関係の理解を支援するインタラクション環境の構築を目指す．本稿では，そのための文脈抽出および可視化の手法について述べ，電子メールを対象に適用例を示し，現時点での課題を明らかにする．

2. 手法

以下，本研究で提案する手法について述べる．提案手法は，文書集合からの文脈の抽出と，抽出された文脈の可視化という 2 つの段階により構成される．以下，それぞれの段階について述べる．

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

2.1 文脈の抽出

時間属性付き文書集合 D を考える．D に含まれる各文書 $d_i \in D$ について，時刻属性 τ_i および，トピック集合 $T_i = \{w_{ij}\}$ が与えられるものとする．ただし，トピック w_{ij} は語彙集合 W の要素とする ($w_{ij} \in W$)．このとき，文書集合 D，トピック w_k についての文脈 c_k は次式で定義される．

$$c_k = \{d_i | \forall i (d_i \in D \wedge w_k \in T_i)\}$$

ここで，各文書についてトピック集合は所与のものとされていることに注意されたい．

2.2 文脈の可視化

文書集合 D 上で定義される文脈集合 $C = \{c_k\}$ について，文脈に含まれる文書 $d_{ki} \in c_k$ に二次元空間上に座標を与え，可視化を行う．文脈 c_k は一定間隔で y 方向に配置される．各文脈に対応する y 座標 y_k とする．各文脈に属する文書は，x 方向については，時刻に対応して位置が決定される．

$$x_{ki} = f(\tau_{ki})$$

ここで，τ_{ki} は d_{ki} の時間属性，f はスケーリング関数である．文脈に属する文書の配置が決定された後，時間的に最古の文書のノードから最新の文書のノードまで線を引く．次に，複数の文脈間で同一の文書を持つ場合には，それぞれの文書ノードを結ぶ形で y 方向に線を引く．可視化の例を図 1 に示す．

3. 電子メールからの文脈抽出

本節では，電子メールを時間属性付き文書とみなし，提案手法を適用した例について述べる．図 1 は提案手法を電子メールの集合に適用して作成した図である．対象としたデータは，人工衛星の開発プロジェクトにおいて，人工衛星に搭載する通信機の仕様についての検討がなされた電子メールのやり取りのなかから 36 通を抜粋したものである．文書の時間属性の期間は 2 月 27 日 16:00 から 3 月 6 日 13:00 の 1 週間である．

トピック抽出の方法として，各文書において語の吸引力に基づいてトピックを抽出する方法を適用した．また，文書毎のトピックだけでなく，文書内に含まれるサブトピックなど，より細粒度でのトピック抽出をおこなう目的で，語の吸引力に基づいた主題遷移分析による文書分割，および文書のトピック構造を表したグラフである WordColony に対する操作による主題

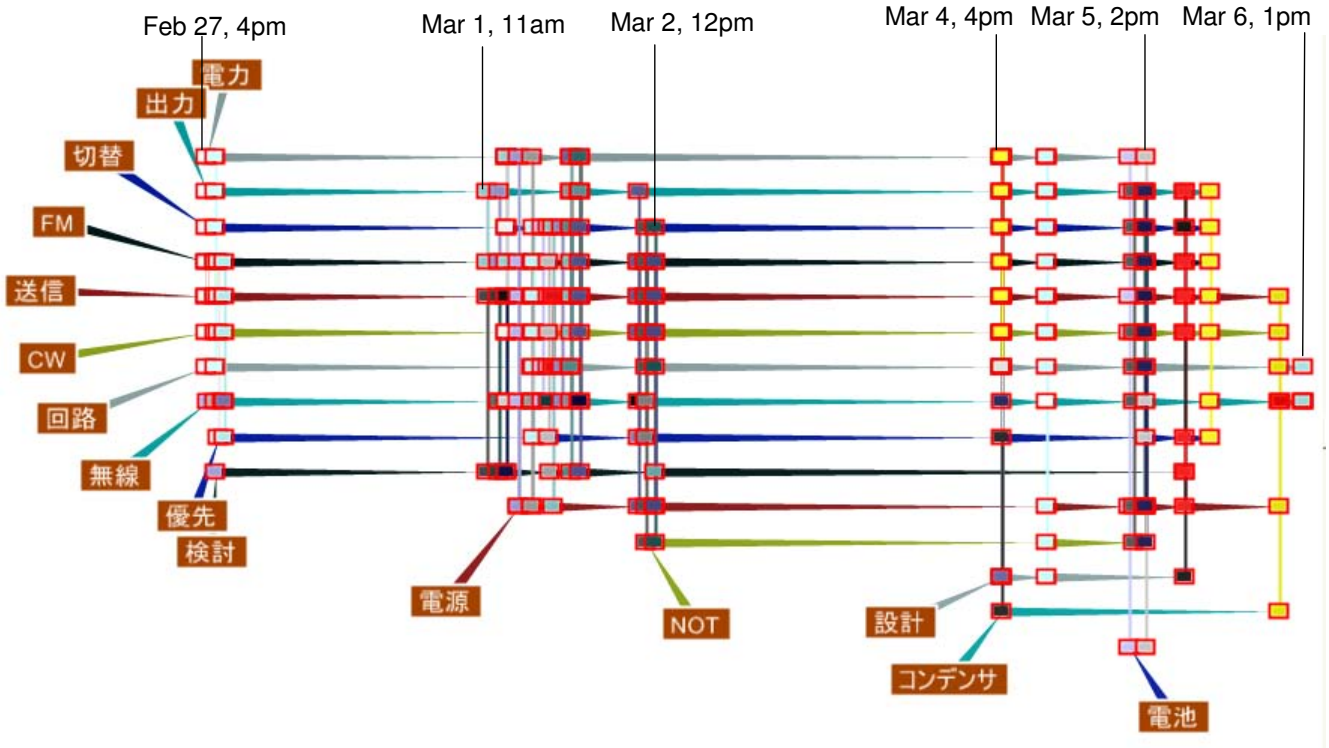

Figure 1: 人工衛星搭載用の通信機の仕様について交わされた 23 通の電子メールに対して，提案手法を適用して作成した図．横軸は時間軸，各行は文脈に対応する

階層分析による文書分割をそれぞれ適用した後にトピックを抽出する方法も適用した．語の吸引力，WordColony，主題遷移分析，主題階層分析についての詳細は [赤石 06] を参照されたい．

図 1 は，各文書から，語の吸引力が最大の語をトピックとして抽出して，提案手法により可視化した結果である．図 2 は主題遷移分析によるトピック抽出による可視化結果を，図 3 は主題階層分析によるトピック抽出による可視化結果をそれぞれ表している．

ある文脈に注目して見ると，期間全体のうち，その文脈のトピックが出現する期間が分かる．例えば，図 1 を見れば，「送信」は期間全般にわたって出現しており，「電源」は期間途中から出現してきていることが分かる．「送信」が期間全般にわたって現れているのは，この文書集合では人工衛星の送信機について議論をしていることが反映されている．また，送信出力との関係で，途中から電源の問題が話題に上ってきているのも，この可視化に反映されている．各文脈上のノードを分布を見ることにより，あるトピックについてどの期間，議論が良くなされたのかが分かる．例えば，「送信」については，期間全体に渡って話題に上っており，「コンデンサ」については，期間後半で少し言及されていることが分かる．このように，提案手法により，文書集合の全体的な傾向と共に，個別の文脈について出現傾向を把握することが可能となる．

可視化結果の中で垂直線（縦糸）で結ばれたノードは同じ文書は全て同一文書を表している．例えば，図 1 において，Mar 4, 4pm 時点で「電力」「出力」「切替」「FM」「送信」「CW」の黄色のノードが縦糸で結ばれているが，この文書において，FM 送信機と CW 送信機で出力の切替器について議論がなされていることを反映したものとなっている．

図 1 と比較して図 2 および図 3 は多くのトピックが現れている．これは，文書分割によりより細粒度のトピックが抽出されているためである．文書毎にトピックを抽出した場合には，全般的な話題を表すような語しか現れず，中で具体的にどのようなことが議論されているのかまでは把握することができなかったが，文書分割適用後のトピック抽出により，より詳細な話題に関連した文脈の可視化ができている．

4. 課題

本稿で示した例は，まだ提案手法の概念実証の段階のものであり，多くの課題を残している．

文脈の配置順序について，現在はユーザの指定した任意の順序で配置される．しかし，文脈間の関連性を何らかの尺度で算出して，関連の近い文脈が近くなるように配置したり，文脈の出現開始時期により並べ替えたりすることにより，より見やすい配置を実現することが可能だと考えられる．

スケーラビリティに関して，文書数および文脈数が多くなるに従って，一画面で全てを表示することが不可能，あるいは一画面で表示しても詳細がよく分からなくなることが考えられる．ズーム機能や縮小表示時の表示方法など，大きなデータについても扱えるよう検討することが必要である．特に文書分割により細粒度のトピックを扱うようになると，表示するトピックの適合率と再現率のトレードオフが問題となる．すなわち，細粒度のトピックを表示することにより，ユーザにとってより関連性の高いトピックが現れる一方（適合率の向上），その他のトピックも多く現れることになる（再現率の低下）．たとえば，トピックの関連性についてクラスタリング手法を適用することにより，ある程度階層的に配置することにより，画面ス

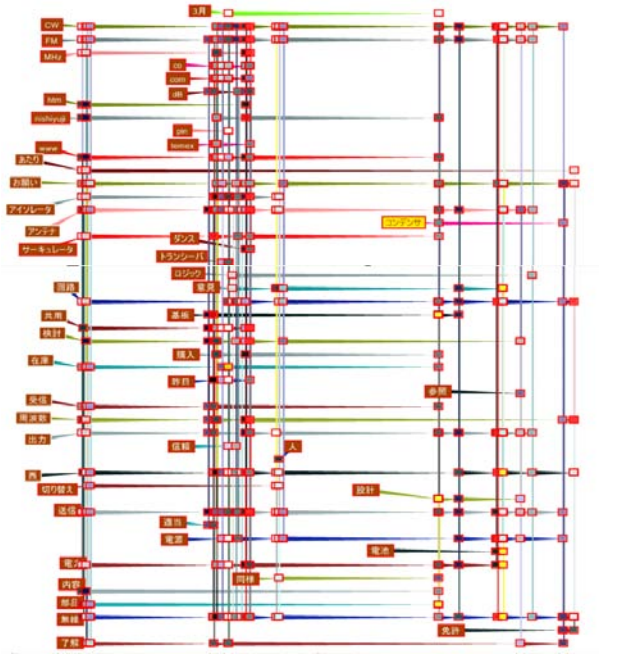

Figure 2: 主題遷移分析により抽出されたトピックを用いた可視化結果

ペースを犠牲にすることなく，より細粒度の文脈を閲覧することが可能にできると考える．

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. 結論

本稿では，時間属性付き文書集合に潜在する多重文脈を抽出し，可視化する手法を提案した．文書集合全体のトピックの傾向を見ると共に，個別のトピックについて文脈として可視化し，更に文脈間の関係も把握できるものである．電子メールを

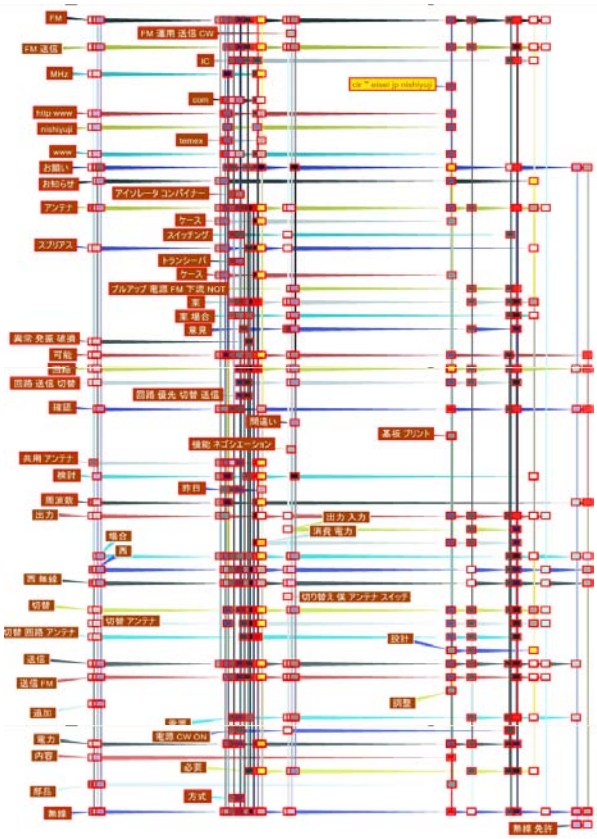

Figure 3: 主題階層分析により抽出されたトピックを用いた可視化結果

対象データとして，提案手法の適用例を示し，現時点での課題を明らかにした．

References

[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)