時間軸を考慮したバネモデルによる文書集合の文脈可視化

Context Visualization of Documents by Incorporating Time Axis into Spring Model

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]東京大学大学院工学系研究科
School of Engineering, The University of Tokyo

With the vast amount of electronic documents available, a way to apprehend the overview of document collections, as well as to understand specific information in context is much needed. In this paper, we propose a method for visuzlizing context by incorporating time axis into spring model. We report a preliminary experiment of visualizing a corpus of design documents from a satellite development project. In particular, we compare the results of temporal-thematic 2D layout and spring model with time axis.

1. はじめに

現代社会は知識化社会と言われており，知識労働者と呼ばれる人たちはアクセス可能な膨大な情報の中から，抱えている問題に関連する情報を収集・分析し，迅速かつ的確にまとめて，自らの，経営層の，あるいは顧客の意志決定に供することが求められている．そのような知識労働者にとって今や必要不可欠となった情報検索技術であるが，ある種の情報要求に対しては十分に応えられていないのが現状である．利用者が，その要求する情報（あるいは検索単位としての文書）が存在することを予め分かっているような場合には有効である（**ナビゲーショナルな情報要求**）．しかし，利用者の要求を満足する情報が存在するかどうか分からない場合，単一の文書では情報要求に応えられず複数の文書の組み合わせが必要な場合，あるいはそもそも利用者自身が情報要求を分節化できていないような場合には必ずしも有効であるとはいえない．後者のような情報要求を満たすには，情報を検索しては解釈し，また検索するといった試行錯誤が必要となる．そのようなタスクを**探索的情報検索**と呼ぶことにする．

知識マネジメントでは，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，文書の時間属性に着目し，文書集合中の時間-主題構造を**文脈**として表現することにより，知識創造につながる探索的情報検索を支援する情報アクセス環境の実現を目指す．

2. 時間-主題構造による文脈の可視化

本章では，時間-主題構造による文脈可視化の枠組みと，文脈可視化を実現する要素技術について述べる．枠組みの概要を 1 に示す．この枠組みでは時間属性付き文書集合 D を対象として，そこから文脈を抽出し可視化する．この枠組みにおいて，**文脈**は時間属性付き基本要素の集合として定義される．**基本要素**とは，あるまとまった単位の文章である．基本要素は

連絡先: 加藤義清，情報通信研究機構，京都府相楽郡精華町光台 3-5，Tel: 0774-98-6874，Fax: 0774-98-6960，E-mail: ykato@nict.go.jp

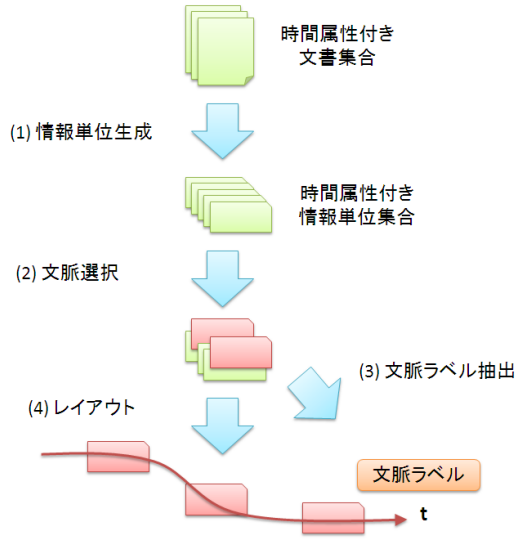

図 1: 時間-主題構造による文脈可視化の枠組み．

D の要素である文書そのものであることもあれば，文書をある基準で分割した**文書断片**であることもある．

2.1 基本要素抽出

基本要素は文脈を構成する基本単位である．基本要素となりうるのは，文書集合中の各文書，何らかの基準に則って文書を分割した結果得られる文書断片，潜在トピック解析により得られる潜在トピックなどである．ここでは，潜在トピックを得る例として潜在ディリクレ割当法 [Blei 03] を取り上げる．

2.1.1 潜在ディリクレ割当法

潜在ディリクレ割当法 (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混合比 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 $p(w|z_i)$ に基づいてテキストが生成されるというものである．

$$\theta_d \sim \mathrm{Dir}(\alpha) \quad (1)$$

$$z_n \sim \mathrm{Multinomial}(\theta_d) \quad (2)$$

$$w_n \sim \mathrm{Multinomial}(\beta_{z_n}) \quad (3)$$

ここで，α はトピック混合比の事前確率を与えるディリクレ分布のパラメータ，θ_d は文書 d のトピック混合比，z_n は θ_d で規定される多項分布に従って選択されたトピック，β_{z_n} トピック z_n の語の多項分布パラメータである．変分ベイズ法などにより近似的にパラメータ推定がなされる．

2.2 文脈選択

文脈選択は時間属性付き基本要素集合 E から何らかの基準でその部分集合 $C \subseteq E$ を選択する操作である．ここでは，文脈選択の方法としてキーワードによる文脈選択，主題による文脈選択，およびクラスタリングによる文脈選択の 3 つの方法について述べる．

2.2.1 キーワードによる文脈選択

キーワードによる文脈選択では，文脈はキーワードで規定され，キーワードが含まれる基本要素が文脈の構成要素として選択される．すなわち，基本要素 e を語の集合として，基本要素集合でキーワード w により規定される文脈 $C(E,w)$ は次のように定義される．

$$C(E,w) = \{e_i | w \in W(e_i) \wedge e_i \in E\} \tag{4}$$

ここで，$W(e_i)$ は基本要素 e_i が含む語の集合である．

2.2.2 主題による文脈選択

主題による文脈選択では，主題を選択することにより，選択された主題を含む基本要素の集合を文脈として扱う．

基本要素に主題を割り当てる方法として，主題遷移解析で用いた吸引力に基づく方法が挙げられる．基本要素が含む語のうち，吸引力が最大の語，あるいは上位 k 個の語をその基本要素の主題を表す語として割り当てるものである．

潜在トピックを基本要素にとる場合，潜在トピックを主題として選択すれば自動的に文脈が決定される．

2.2.3 クラスタによる文脈選択

クラスタリングによる文脈選択では，基本要素にクラスタリング法を適用して得られるクラスタが文脈を形成しているとみなし，クラスタを選択することにより文脈を選択する．クラスタリング法を適用するためには基本要素間で非類似度を定義する必要がある．非類似度を与え方としては，基本要素に含まれる語に基づくベクトル空間法や，LDA などの多重潜在トピック分析を適用して基本要素のトピック比に基づくベクトル空間での距離を利用することができる．

2.3 文脈ラベル抽出

文脈を可視化する際，文脈が何を表しているのかを示すラベルを付与することは，ユーザの理解にとって重要である．キーワードや主題語に基づく文脈選択の場合は，キーワードや主題語をそのままラベルとして利用できるが，潜在トピックやクラスタによる文脈選択においては何をラベルとして表示するかが問題となる．

2.4 レイアウト

レイアウトの段階では，文脈選択により得られる文脈は基本要素の時間属性，非類似度などに基づき二次元平面上に配置する．

2.4.1 時間-主題軸による二次元配置

基本要素集合 $E = \{e_i\}$ 上で定義される文脈 $C_k \subset E$ について，文脈の要素 $e_i^k \in C_k$ 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 軸方向に配置される．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．

$$x_{ki} = \xi\tau(e_i^k) \tag{5}$$

ここで，$\tau(e_i^k)$ は e_i^k の時間属性，ξ はスケーリング係数である．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．

基本要素が潜在トピックである場合，基本要素は属性としてトピック比に基づくトピックの強さを持つ．この場合，トピックの強さをノードの大きさで表現する．

2.4.2 時間軸拘束バネモデルに基づくレイアウト

時間軸拘束バネモデルに基づくレイアウトでははは，2 次元平面上で，1 つの軸を時間軸，他方の軸を主題軸として選び，基本要素を配置する．基本要素はその時間属性の値に応じて，時間軸上に配置される．主題軸上の座標はは同一時間区分内の他の基本要素，および隣接する時間区分内の基本要素とをバネで接続したバネモデルの最適化により与える．このとき，基本要素の時間軸座標は初期値で固定する．バネの自然長は基本要素間の被類似度に比例する形で与える．以下に，時間軸拘束バネモデルに基づくレイアウト法の手順を示す．

1. 時間軸を一定の区間（1 日，1ヶ月など）毎に区分し，各区分に対して時間軸上の代表値を与える．
2. 各基本要素をその時間属性に基づき，先に定義した時間区分に割り当てる．このとき，割り当てられた時間区分の代表値が，その基本要素の時間軸上の初期座標となる．
3. 同一時間区分に属する基本要素間，および隣接時間区分の基本要素間をバネで接続する．バネの自然長は基本要素間で定義される被類似度に基づいて与える．
4. 式 (6) バネモデルで定義される系のエネルギーが最小となるよう，基本要素座標 $\mathbf{x}_i = (t_i, x_i)$ の x_i について最適化する．

以下に，各基本要素の座標を (t_i, y_i) として時の時間拘束バネ・マスモデルの系のエネルギーおよび，エネルギーの主題軸 (y_i) についての微分の式を与える．

$$\mathcal{E}(y_1, ..., y_n) = \sum_{s \in S} \frac{1}{2} k(l_s - \bar{l}_s)^2 \tag{6}$$

$$l_s = \sqrt{(t_{s1} - t_{s2})^2 + (y_{s1} - y_{s2})^2} \tag{7}$$

$$\frac{\partial \mathcal{E}}{\partial y_i} = \sum_{s \in S(y_i)} \frac{1}{\sqrt{l_s}} k(l_s - \bar{l}_s)(y_{s1} - y_{s2}) \tag{8}$$

ここで，S は系のバネの集合で，$s = (e_i, e_j) \in S$ はバネ s が接続するノードの対である．l_s および $\bar{l}_s$，それぞれバネ s の長さおよび自然長を，$S(e_i) = \{s | e_i \in s \wedge s \in S\}$ はノード e_i が接続するバネの集合を表す．

3. 実験

時間軸拘束バネモデルによる文書集合可視化の効果を評価するために，同じ文書集合を時間-主題軸に基づく二次元配置と時間軸拘束バネモデルという 2 つのレイアウト法で可視化し，比較をおこなった．

3.1 データ

実験に用いたデータは位置天文観測小型衛星を開発する東京大学中須賀研究室 Nano Jasmine プロジェクト[*1] より提供さ

*1 http://www.space.t.u-tokyo.ac.jp/nanojasmine/Index.htm

れた 889 件の設計文書である．設計会議の資料，議事録，設計計算書などが含まれる．

3.2 方法

以下の方法で，可視化を行った．まず，文書ファイルからテキストを抽出し，MeCab[*2]を用いて形態素解析を施した．得られた形態素列から文書集合中で 10 回以上出現する名詞のみを抜き出し，更に記号や数字のみからなる形態素，アルファベットやひらがな 1 文字のみからなる形態素などをストップワードとして除いた結果，2348 語からなる語彙集合を得た．この語彙集合に基づいて，文書集合に対してトピック数 100 で LDA 法を適用し，結果として得られる変分パラメータ γ_i を $\sum \gamma_i = 1$ となるように正規化したものを文書ベクトルとした．

得られた文書ベクトルに基づき，時間-主題軸に基づく二次元配置および時間軸拘束バネモデルによるレイアウト法を適用し，可視化を行った．

3.3 結果と考察

図 2 に時間-主題軸に基づく二次元配置による文書集合可視化の結果を示す．この結果から，時間的に局在するトピックが捉えられていることが分かる．時間的に局在するトピックについて，トピックに対応する γ が高い文書を精査した，その内容を表すラベルを図中に示す．精査の結果，LDA によって獲得されたトピックに 2 種類のものがあることが判明した．1 つは黄緑色のラベルがついているトピックで，設計開発の内容（「熱構造モデル試験」「シミュレータ」「アンテナ」など）に対応するトピックである．もう 1 つは，文書のタイプあるいは開発時期に対応するトピックである．これは，主に特定の時期の進捗報告あるいは議事録が同一のトピックに高い γ の割当を受けているものである．前者は LDA によって獲得が期待されたトピックであるが，後者のようなトピックは予期していなかった．後者のようなトピックが獲得される理由として，開発のある一定期間同じ種類の検討，設計，あるいは試験が並行して続くことが多く，それらについて記述される進捗報告や議事録を同じトピックに属すると判断するからと考えられる．

次に，図 3 に，時間軸拘束バネモデルに基づく文書集合可視化例の結果を示す．この可視化においては，k-means 法により 100 クラスタまでまとめた後，Ward 法による階層クラスタリングで，階層クラスタを得た．図はクラスタ数が 10 となるように，クラスタを選んで可視化したものである．この結果から，二次元配置による可視化法では分からないクラスタ間の距離の時間変化を可視化することが可能であることを確認できた．本実験では，特定のクラスタに多くの文書が所属する結果となって，それぞれのクラスタが捉える文脈が必ずしも明らかではなかった．有効な文脈を捉えたクラスタを得るためのクラスタリング手法の検討は今後の課題である．

4. 関連研究

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．

PaperLens [Lee 05] は論文を対象とした可視化システムである．与えられた論文集合におけるトピックのトレンドを，トピックごとに各年における論文数を棒グラフとして表示でき，ユーザが指定する著者や引用関係により，条件に該当する論文がグラフ中でハイライトされるようになっていて，ある著者がどの時期にどの分野で論文を書いたのか，ある論文がどのような分野でよく引用されているのか，などといった情報要求に応えられるようになっている．

本研究では，統計的トピックモデルである LDA を文脈抽出手法に応用しているが，関連する研究として統計的トピックモデルを可視化に応用したもの [Iwata 08] や統計的なトピックモデルをトピックの時系列変化の解析に応用するもの [Griffiths 04, Blei 06, Li 06, Wang 06, Wang 08] が挙げられる．Griffiths ら [Griffiths 04] は 1991 年から 2001 年の米国科学アカデミー紀要 (PNAS) で出版された論文の要約に LDA を適用し，得られた要約のトピック混合比 θ を年ごとに平均したものの変化から，平均トピック混合比が上昇を続けているものをホットトピック，下降を続けているものをコールドトピックとして分析している．これは，トピックの時間変化を考慮に入れない LDA を適用した後，事後的に文書集合中のトピックの時間的

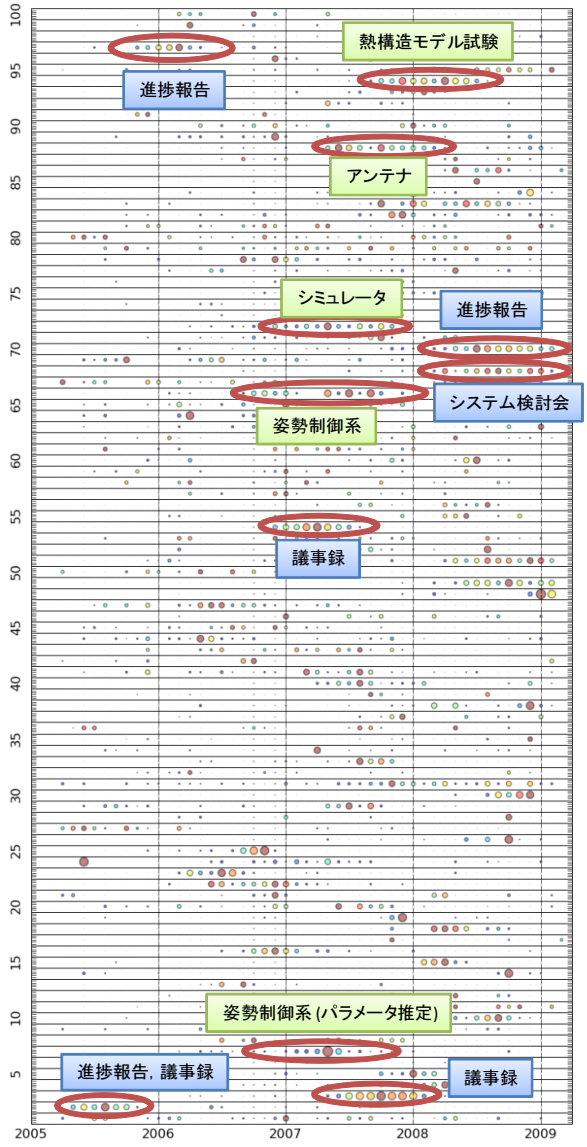

図 2: 時間-主題軸に基づく二次元配置による文書集合可視化.

*2 http://mecab.sourceforge.net/

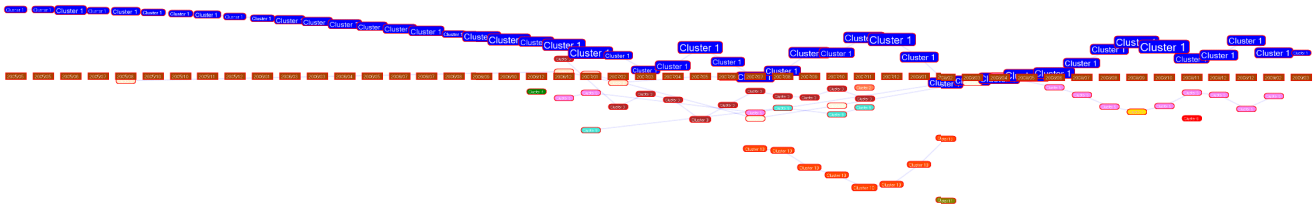

図 3: 時間軸拘束バネモデルによる文書集合可視化例.

変化を分析したものである.

トピックの時間変化を考慮したトピックモデルとしては，時間を離散的に扱ったものと連続的に扱ったものに分類できる．前者の例として，Blei ら [Blei 06] による，LDA を拡張した動的トピックモデル (DTM) が挙げられる．DTM ではトピック比およびトピック毎の語の出現確率分布を決定するパラメータによる状態空間がマルコフ過程により変化をする．後者の例としては，Wang ら [Wang 06] による Topic over Time (TOT) が挙げられる．TOT では，トピックに連続時間軸上での分布を仮定することにより，出現語彙が共通する時間的には離れたトピックも区別して扱うことができる．

5. おわりに

本論文では，文書集合の文脈可視化の方法として，時間軸拘束バネモデルによる可視化手法を提案し，時間-主題構造に基づく二次元配置との比較をおこなった．時間軸拘束バネモデルによりクラスタ間の距離の変化を可視化することが可能であることが確認できた．有効な文脈を捉えるためのクラスタリング手法の開発は今後の課題である．

謝辞

本研究に使用したデータは東京大学中須賀研究室 Nano Jasmine プロジェクトより提供を受けました．ご協力いただいた同プロジェクトのプロジェクトマネージャ酒匂信匡助教に感謝いたします.

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)

[Blei 06] Blei, D. M. and Lafferty, J. D.: Dynamic Topic Models, in *Proceedings of the 23 rd International Conference on Machine Learning* (2006)

[Furnas 86] Furnas, G. W.: Generalized fisheye views, in *CHI '86: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 16–23, New York, NY, USA (1986), ACM

[Griffiths 04] Griffiths, T. L. and Steyvers, M.: Finding scientific topics, *Proceedings of the National Academy of Sciences of the USA*, Vol. 101, pp. 5228–5235 (2004), suppl. 1

[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)

[Iwata 08] Iwata, T., Yamada, T., and Ueda, N.: Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents, in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, pp. 363–371 (2008)

[Lee 05] Lee, B., Czerwinski, M., Robertson, G., and Bederson, B. B.: Understanding research trends in conferences using PaperLens, in *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pp. 1969–1972, New York, NY, USA (2005), ACM

[Li 06] Li, W., Wang, X., and McCallum, A.: A Continous Time Model of Topic Co-occurrence Trends, in Ashish, N., Appelt, D., Freitag, D., and Zelenko, D. eds., *Event Extraction and Synthesis: Papers from the 2006 AAAI Workshop*, pp. 48–53, AAAI Press (2006), [Technical Report, WS-06-07]

[Mackinlay 91] Mackinlay, J. D., Robertson, G. G., and Card, S. K.: The perspective wall: detail and context smoothly integrated, in *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, pp. 173–176, New York (1991), ACM Press

[Wang 06] Wang, X. and McCallum, A.: Topics over time: a non-Markov continuous-time model of topical trends, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, pp. 424–433 (2006)

[Wang 08] Wang, C., Blei, D., and Heckerman, D.: Continuous time dynamic topic models, in *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)* (2008)

[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)