1H1-1

SalienceGraph: 参照確率に基づく話題遷移図の可視化

SalienceGraph: Visualization of Topic Transition Based on Reference Probability

白松 俊[*1*2] Shun SHIRAMATSU　久保田 祐史[*1] Yuji KUBOTA　駒谷 和範[*1] Kazunori KOMATANI　尾形 哲也[*1] Tetsuya OGATA　奥乃 博[*1] Hiroshi G. OKUNO

[*1]京都大学 情報学研究科 Graduate School of Informatics, Kyoto University

[*2]日本学術振興会特別研究員 PD Research Fellow (PD), Japan Society for the Promotion of Science

Discourse context dynamically changes because the targets of the participants' attention change with each utterance unit. This paper presents visualization schemes of the dynamic flow of topics in discourse context. Visualizing the topic dynamics in a long discourse helps its readers to understand its time-series contextual overview. To visualize the topic dynamics, we define reference probability as a quantitative scale of discourse salience on the basis of centering theory. A dynamic context can be represented as a high dimensional vector of the reference probability. To visualize it in three dimensional space, we reduce dimensions by using PLSA (Probabilistic Latent Semantic Analysis). Words associated with latent topics on the basis of weighted PMI (Pointwise Mutual Information) help users to grasp the topics. Our visualization will enable us to develop browsing interface for long meeting minutes, which satisfies Shneiderman's "Visual information-seeking mantra".

1. はじめに

背景と目的　談話文脈中の話題は発話ごとに動的に変化し，談話全体で「流れ」を形成する．本稿の目的は，話題の流れが一目で把握できるような話題遷移図の可視化である．このような話題遷移図は，例えば議事録などの長い談話を閲覧する際に役に立つ．全てを読まなくても視覚的に話題の流れを把握できれば，談話全体の流れの中から所望の議論を発見しやすくなるからである．つまり，Shneiderman が提唱した "Visual information-seeking mantra"，すなわち「Overview first, zoom and filter, then details on demand (まず最初に概観を見せ，ズームやフィルタ操作の後，要求に応じて詳細化を行う)」[Shneiderman 98] という要件を満たす談話閲覧インタフェースの実現に繋がると期待される．

アプローチ　この目的のためにわれわれは，各発話ごとに単語やトピックへの注目度を計測するというアプローチをとる．単語やトピックへの注目度を，本稿では顕現性 (salience) と呼ぶ．顕現性を計測するための尺度としては，過去の研究においてわれわれが設計した参照確率 (reference probability) を用いる [白松 07]．これにより，発話系列 $[U_1, \ldots, U_n]$ 中の任意の発話 U_i における文脈を，その瞬間の各単語の参照確率から成るベクトルで表すことができる．しかしこのベクトルは高次元なので，可視化に適した少数のトピックに縮退させるため，PLSA (Probabilistic Latent Semantic Analysis) を用いる．

このようなアプローチにより，例えば図 1 のような話題遷移図を描くことができる．われわれは，このような話題遷移図を **SalienceGraph** と名付ける．

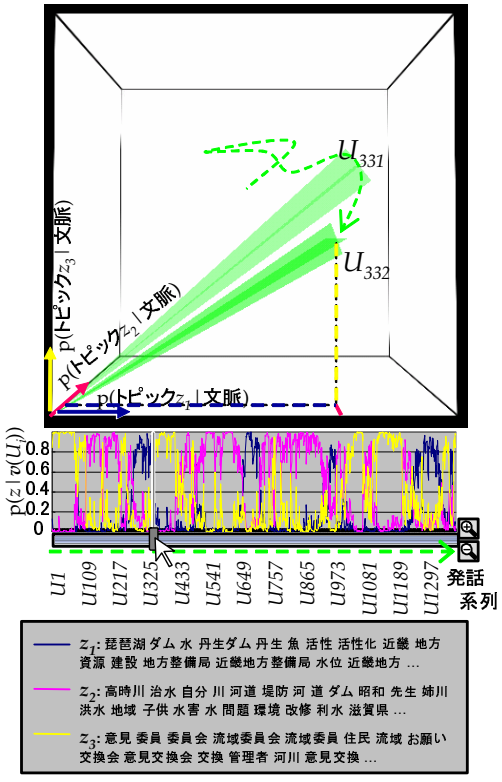

図 1: SalienceGraph の例

2. 参照確率: 動的に変化する顕現性の計算

われわれは過去の研究において，談話中に現れる単語 w の顕現性を以下のように定義した．

$$\begin{aligned}(U_i\text{での } w \text{ の顕現性}) &= p(\exists w' \xrightarrow{\text{coref}} w \text{ in } U_{i+1}|\text{pre}(U_i)) \\ &= p(w|\text{pre}(U_i))\end{aligned}$$

ただし，U_{i+1} は後続発話を，$\text{pre}(U_i)$ は先行文脈 $[U_1, \ldots, U_i]$ を表し，$w' \xrightarrow{\text{coref}} w$ in U_{i+1} は発話 U_{i+1} 中の単語 w' が w と共参照関係にあることを表す．この定義は，「目立っている実体ほど次の発話で参照される可能性が高い」というセンタリング理論 [Grosz 95] の知見に基づいており，センタリング理論との整合性が確認されている [白松 07]．$p(w|\text{pre}(U_i))$ の計算は，基本的には以下の式のように訓練コーパス中のサンプル $\langle w_c, U_j \rangle$ を用いて行う．

連絡先: 白松 俊, 〒 606-8501 京都市左京区吉田本町 京都大学 情報学研究科 奥乃研究室, siramatu@kuis.kyoto-u.ac.jp

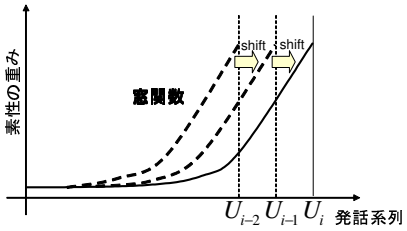

図 2: 1 発話毎にシフトする窓関数で新近性効果を扱う

$$p(w|\mathrm{pre}(U_i)) = \frac{\#\{\langle w_c, U_j\rangle; \mathrm{feat}(w_c, U_j) = \mathrm{feat}(w, U_i) \wedge \exists w_c' \xrightarrow{\mathrm{coref}} w_c \text{ in } U_{j+1}\}}{\#\{\langle w_c, U_j\rangle; \mathrm{feat}(w_c, U_j) = \mathrm{feat}(w, U_i)\}}$$

ここで $\mathrm{feat}(w_c, U_j) = \mathrm{feat}(w, U_i)$ は，対象事例 $\langle w, U_i\rangle$ の素性ベクトルと，訓練コーパス中のサンプル $\langle w_c, U_j\rangle$ の素性ベクトルとが等しいことを表す．

ただし，実際の計算では上記の式のようにサンプルの個数を数えるのではなく，データスパースネスを避けるため，あらかじめ訓練コーパスから学習した回帰モデル (e.g., ロジスティック回帰) を用いる．計算のための素性としては，新近性[*1](w の出現箇所と U_{i+1} の発話距離)，文法役割，品詞，出現頻度を用いる．新近性効果の減衰曲線を図 2, 3 のような窓関数によって扱い，図 4 のようにロジスティック回帰によって $\mathrm{pre}(U_i)$ を計算する．

素性選択や窓関数の選択などの最適化基準としては，やはりセンタリング理論の知見を参考にして,「後続発話における参照の予測能力」を表す以下のような評価尺度を設計した [白松 08].

ある顕現性の推定手法 m の評価尺度 $\mathrm{evalSal}(m)$ を，以下のように定義する．

$$\mathrm{evalSal}(m) = \mathrm{cor}\Big(\big[\mathrm{sal}_m(w|\mathrm{pre}(U_i))\big]_{\langle w,U_i\rangle}, \big[\mathrm{isRef}(w, U_{i+1})\big]_{\langle w,U_i\rangle}\Big)$$

ただし，$\mathrm{cor}(\cdot,\cdot)$ は相関係数，$\langle w, U_i\rangle$ は実体 w と発話 U_i の組から成るサンプル，$\mathrm{sal}_m(w|\mathrm{pre}(U_i))$ は m により計算される U_i における w の顕現性の値，$\mathrm{isRef}(w, U_{i+1})$ は後続発話における w への参照の正解 $\mathrm{isRef}(w, U_{i+1}) = \big(1 \text{ if } (\exists w' \xrightarrow{\mathrm{coref}} w \text{ in } U_{i+1}),\ \text{otherwise } 0\big)$ である．この $\mathrm{evalSal}(m)$ は，テストセットコーパスを用いて計測され，後続発話における参照の予測性能を表す．

この評価尺度を用い，図 4 の計算手法を最適化した．更に評価実験として，新聞記事コーパス (毎日新聞 3,000 記事) と話し言葉コーパス (CSJ; 日本語話し言葉コーパスのインタビュー 4 対話) の上で，単純なベースライン手法 (矩形窓を区切って数えた単語の出現頻度) と比較した．その結果，両コーパスで後続発話における参照の予測能力が向上し，特に，話し言葉コーパスで大幅な向上を確認した [白松 08]．このことから，図 4 の顕現性計算手法は**特に話し言葉で有効である**という知見を得た．

ここまで本節では，われわれが以前開発した参照確率計算手法の概要を述べた．以後の節ではこの手法を発展させ，話題遷移の可視化に適した手法へと拡張する．

*1 新近性効果 (recency effect) とは,「最近の出来事ほど思い出されやすい」という認知科学上の知見である．

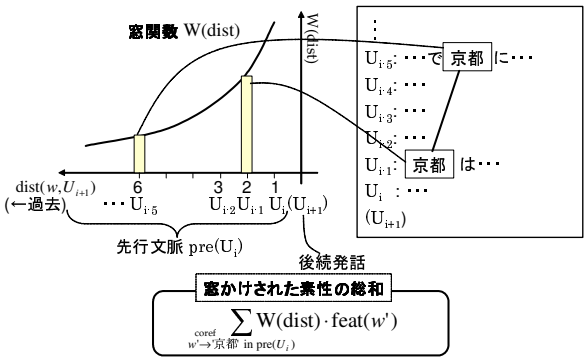

図 3: 窓関数を用いた素性の重み付け

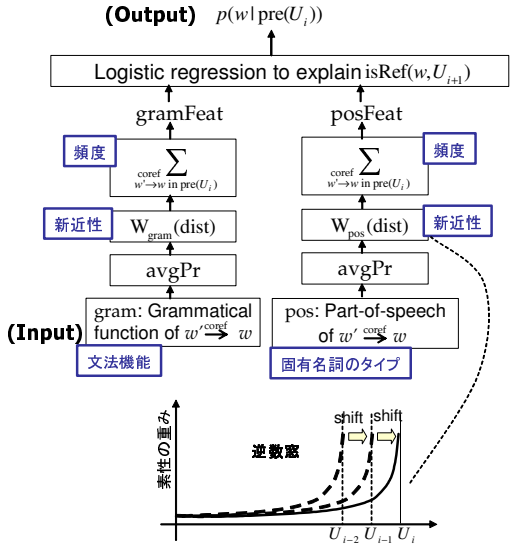

図 4: 逆数窓関数とロジスティック回帰による参照確率の計算

3. 次元圧縮: 参照確率と PLSA の統合

可視化の対象となる談話中に N 種類の名詞句 $w_1, \cdots, w_N$ が含まれているとき，談話中の任意の発話 U_i の瞬間の文脈は，以下のような参照確率から成るベクトルによって表すことができる．

$$\boldsymbol{v}(U_i) = \big[p(w_1|\mathrm{pre}(U_i)), \cdots, p(w_N|\mathrm{pre}(U_i))\big]^T$$

しかし，この N 次元の参照確率ベクトルは高次元なので可視化には適していないという問題点がある．この問題の解決策としては，以下の 2 つが考えられる．

解決策 (a) 参照確率ベクトルを少数の潜在トピック次元に圧縮し，その潜在トピックの生起確率を可視化に用いる．

解決策 (b) 少数の重要語を抽出し，その参照確率を可視化に用いる．

本稿では，解決策 (a), (b) を併用して話題遷移を可視化するアプローチをとる．すなわち，談話閲覧インタフェースにおいて，まず “Overview first” の段階で (a) を用い，後から “zoom and filter” の操作を行うためにユーザが (b) を利用する．(a) のためには，参照確率と PLSA (Probabilistic Latent

Semantic Analysis) [Hofmann 99] を統合する．(b) のためには，PLSA によって決定した潜在トピック z との重み付き PMI (Pointwise Mutual Information; 自己相互情報量)，すなわち $p(w|z)PMI(w,z) = p(w|z)\log\frac{p(w,z)}{p(w)p(z)}$ が上位になった単語 w のリストを重要語群と見なし，その中からユーザが所望の単語を選択できるようにする．

3.1 参照確率ベクトル集合からの PLSA の学習

PLSA のモデル学習のための訓練データとして，対象とする談話 $[U_1,\ldots,U_n]$ の参照確率ベクトル集合 $\{\boldsymbol{v}(U_i)|i = 1,\ldots,n\}$ を用いる．つまり，一般的な PLSA の学習では文書 d や矩形窓で区切られた履歴 h における単語出現頻度から成る Bag-of-Words ベクトルの集合を訓練データとして用いるが，本稿ではそれらの代わりに参照確率ベクトルの集合を用いることで，参照確率と PLSA を統合する．モデル学習においては，EM アルゴリズムで以下の対数尤度 L を最大化するようなパラメタ $p(z_k)$, $p(w_j|z_k)$, $p(\boldsymbol{v}(U_i)|z_k)$ を訓練データから学習する．

$$\begin{aligned} L &= \sum_{i,j} p(w_j|\mathrm{pre}(U_i))\log p(w_j,\boldsymbol{v}(U_i)) \\ &= \sum_{i,j} p(w_j|\mathrm{pre}(U_i))\log\sum_k p(z_k)p(w_j|z_k)p(\boldsymbol{v}(U_i)|z_k) \end{aligned}$$

本稿では，潜在トピック z の数は 3 に設定する．理由は，図 1 上部のような 3 次元表示をするためである．また，z が 4 つ以上になると図 1 下部の話題遷移図も混み合い，図の可読性が低下する恐れがある．

3.2 重みつき PMI による重要語抽出

本稿では潜在トピック z に対し，重みつき PMI

$$p(w|z)PMI(w,z) = p(w|z)\log\frac{p(w,z)}{p(w)p(z)}$$

が大きな単語 w のリストを関連づけて表示する．PMI は，共起の度合として使われることが多い尺度である．ここで $PMI(w,z)$ をそのまま使うのではなく $p(w|z)$ で重み付けするのは，話題遷移の可視化には適さない低頻度語が上位に来るのを避けるためである．

4. 長い議事録への適用

淀川水系流域委員会が Web サイトで公開している議事録 [淀川水系流域委員会 05] を用い，実際に議事録を可視化する例を示す．これは，ダムに関する議論 1,394 文から成る長い議事録である．以下では 1 文を 1 発話単位と見なした．また参照確率 $p(w|\mathrm{pre}(U_i))$ の計算については，CSJ (日本語話し言葉コーパス) のインタビュー 4 対話から獲得したロジスティック回帰モデル [白松 08] を用いて行った．

表 1 に，各潜在トピック z_1, z_2, z_3 に対して $p(w|z)PMI(w,z)$ が上位であった単語群を示す．議事録の内容と照らし合わせると，z_1 は地域振興や地域の活性化を望む住民の意見に対応し，z_2 は治水や洪水被害に関する意見に対応し，z_3 は司会や議事進行の発話に対応している．

図 5 に，議事録閲覧時の操作手順の例を示す．(1) まず z_1, z_2, z_3 とその関連語を用いて自動的に概観を表示し，(2) 次に，それを見たユーザが興味に応じて z_3 を外し，語「治水」を追加する．(3) z_1, z_2,「治水」が混ざっていそうな U_{284} 付近を拡大し，(4) そこをダブルクリックして議事録を表示する．

表 1: 参照確率から推定した潜在トピックを代表する単語群

z	$p(w\|z)PMI(w,z)$ が上位の単語 w	解釈
z_1	琵琶湖 ダム 水 丹生ダム 丹生 魚 活性 活性化 近畿 地方 資源 建設 地方整備局 近畿地方整備局 水位 近畿地方 周辺 平成 話 国 人たち 流下 事業 開発 目的 橋本先生 地方整備 近畿地方整備 瀬切れ 水資源 余呉町 ダム建設 ダム周辺 泥水 流量 京都 大阪 全国 動植物 方針案 治水や 地下水 瀬 本体	地域振興や活性化など
z_2	高時川 治水 自分 川 河道 堤防 河 道 ダム 昭和 先生 姉川 洪水 地域 子供 水害 水 問題 環境 改修 利水 滋賀県 家 用水 びわ もん 被害 滋賀 矢板 虎姫 びわ町 底 道改修 河道改修 方法 全部 生活 さかい 視点 金 盛 伏流水 集落 災害 伐採 工 頭首工	治水や災害被害など
z_3	意見 委員 委員会 流域委員会 流域委員 住民 流域 お願い 交換会 意見交換会 交換 管理者 河川 意見交換 河川管理者 ご意見 管理 河川管理 発表 委員長 淀川 紹介 皆さん 発表者 一般 計画 傍聴者 反映 傍聴 方々 討論 発言 整備 水系流域委員会 淀川水系流域委員会 ご紹介 学識経験者 経験者 反対説明 一般傍聴者 形 中 水系流域委員 水系流域	司会や議事進行など

本稿で開発した要素技術により，このような “Visual information-seeking mantra” を満たす談話閲覧インタフェースが実現可能になったので，今後はこれに基づいて談話閲覧システムを実装・開発する予定である．

5. 関連研究

5.1 長距離言語モデルによる文脈推定

文脈を捉えるための PLSA に基づく長距離言語モデルの研究では，訓練データには矩形窓で区切られた履歴 h を表す Bag-of-Words ベクトルを用いるのが一般的である．これに対してわれわれの手法では，訓練データに参照確率ベクトルを用いている点が異なる．参照確率ベクトルは，矩形窓で区切られた頻度に基づく Bag-of-Words に対し，後続発話における参照の予測性能が上回ることが明らかになっている [白松 08] が，PLSA と統合した場合の性能比較は行っていない．これは今後の課題としたい．

持橋らは，長距離言語モデルに用いられる LDA (Latent Dirichlet Allocation) や DM (Dirichlet Mixture) をパーティクルフィルタと組み合わせた MSM-LDA (Mean Shift Model-LDA), MSM-DM (Mean Shift Model-DM) を提案している [持橋 05]．これは，履歴 h の窓の長さを固定せず最適な文脈長を自動的に選択する手法であり，SalienceGraph の可視化にも適している可能性がある．今後，これらの手法の導入も検討したい．

5.2 年代順の話題遷移図

本研究の SalienceGraph は 1 つの談話を発話や文の時系列的ストリームと見なして話題遷移を可視化する手法である．一方，文書集合を文章の年代順ストリームと見なして話題遷移を可視化する手法が盛んに研究されており，例えば Kleinberg の word burst [Kleinberg 02] や，blog 界の話題遷移を可視化する kizasi というサイト*2などが有名である．これらの年代・月日順の話題遷移図では，ストリームの単位が文書という大きな単位であるので，単語の出現頻度が有効である．それに対し，本研究の SalienceGraph ではストリームの単位が発話や

*2 http://kizasi.jp/

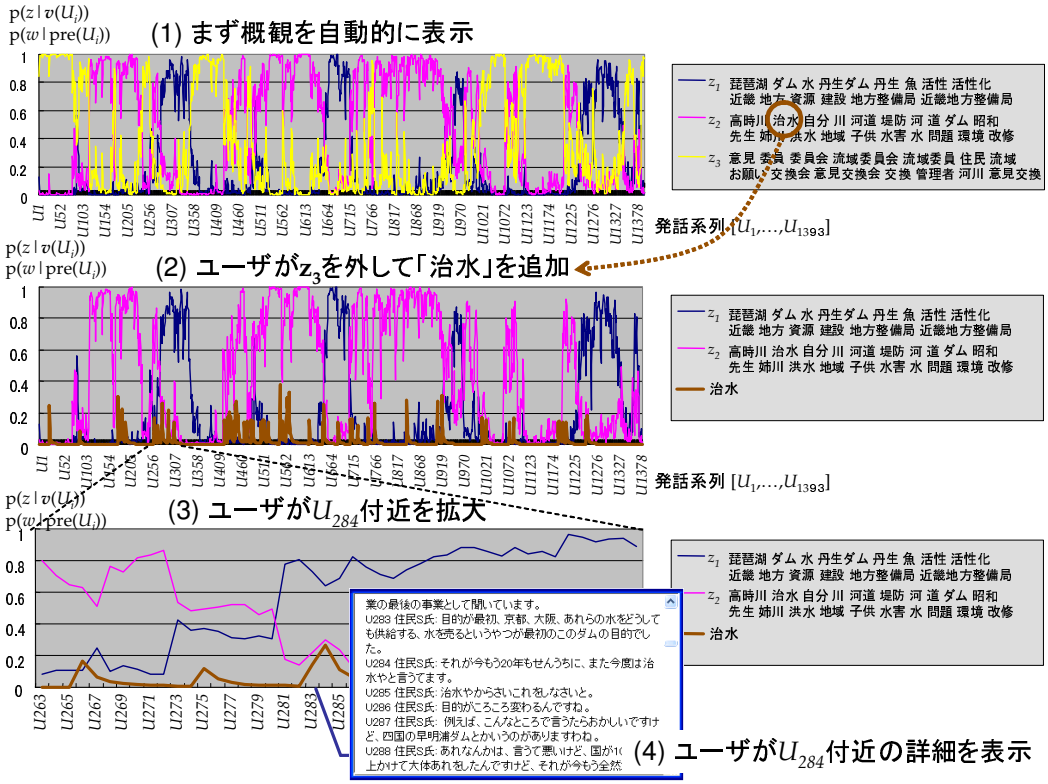

図 5: 話題遷移図 SalienceGraph を用いた議事録閲覧の例

文という小さな単位であるため，単純な出現頻度ではなく，参照確率と PLSA を統合した手法を開発する必要があった．

6. まとめ

本稿では，**一目で長い談話の「流れ」を把握できる可視化手法 SalienceGraph** を開発した．具体的には，後続発話の予測能力に優れた参照確率を，確率的言語モデルに用いられる PLSA と統合することにより，Shneiderman の “Visual information-seeking mantra” を満たす談話閲覧インタフェースのために必要な要素技術を開発した．これにより，以下の機能が実現可能となった．

(A) 潜在トピックを用いた “Overview first”

(B) 重みつき PMI の値で潜在トピックに関連付けられた語を用いた “zoom and filter” 操作

(C) SalienceGraph 上の任意の箇所を指定して議事録を表示する “detail on demand” 操作

(D) 3 次元空間上のベクトルの動きとしての話題遷移の可視化

現在までに，(D), すなわち図 1 上部の 3 次元表示部分は実装済みである．その際，音環境の可視化のために開発された SmartRoomViewer [吉田 07] のソースコードを一部流用した．今後は，談話閲覧システム全体の完成を目指す．開発したシステムは，公的討議の分析の研究に利用する予定である．

謝辞 公的討議の分析という観点から多くの示唆を頂いた鄭蝦榮さん，小林潔司先生に感謝致します．本研究は，科研費 (特別研究員奨励費) の支援を受けて遂行されました．

参考文献

[Shneiderman 98] B. Shneiderman: *Designing the User Interface Strategies for Eective Human-Computer Interaction*, 3rd edition, Addison-Wesley, Chapter 15 (1998).

[白松 07] 白松俊, 駒谷和範, 橋田浩一, 尾形哲也, 奥乃博: ゲーム理論に基づく参照結束性のモデル化と日本語・英語の大規模コーパスを用いた統計的検証. 言語処理学会 自然言語処理 14(5), pp.199-239 (2007).

[Grosz 95] B. Grosz and A. Joshi and S. Weinstein: Centering: A Framework for Modeling the Local Coherence of Discourse, *Computational Linguistics*, 21(2), pp.203-226 (1995).

[白松 08] 白松俊, 駒谷和範, 尾形哲也, 奥乃博: 新近性効果の減衰曲線を加味した顕現性計算手法に基づく話題遷移の可視化. 言語処理学会 第 14 回年次大会発表論文集, pp.432-435 (2008).

[Hofmann 99] T. Hofmann: Probabilistic Latent Semantic Indexing, In Proc. of *SIGIR '99*, ACM Press, pp.50-57 (1999).

[淀川水系流域委員会 05] 淀川水系流域委員会: 住民と委員との意見交換会（丹生ダム）議事録, http://www.yodoriver.org/kaigi/biwa/h17.html#ikenkoukan (2005).

[持橋 05] 持橋大地, 松本裕治: Particle Filter による文脈の動的ベイズ推定. 情報処理学会研究報告 2005-NL-165, pp.59-66 (2005).

[Kleinberg 02] J. Kleinberg: Bursty and Hierarchical Structure in Streams. In Proc. of *the 8th ACM SIGKDD International Conference on Knowledge Discovery and DataMining*, pp. 1-25 (2002).

[吉田 07] 吉田雅敏, 海尻聡, 山本俊一, 中臺一博, 駒谷和範, 尾形哲也, 奥乃博: 音環境を可視化する録音再生システム. 情報処理学会 第 69 回全国大会, 2N-6 (2007).