論文ゼミ#1

戸田浩之, 北川博之, 藤村考, 片岡良治, 奥雅博,2007,
「グラフ分析を利用した文書集合からの話題構造マイニング」

修士2年 松村 草也

研究の背景

- ✓ 人口減少時代において，地方都市，もしくはさらにルーラルな地域における持続的なコミュニティの維持において，近年の情報技術が寄与することの如何に本研究の目的はおかれる.
- ✓ たとえば地域SNSのようなツールが個人，NPOなどを主体とした任意団体によって運営されているが，そのシステムの網羅性はまだまだ弱く，インターネットの利点が有効に生かせているとは考えにくい.
- ✓ しかし一方でそのようなシステムにすべての管理をゆだねることは，人と人のリアルなコミュニケーションから離れ，真の地域コミュニティ維持につながるかといえば議論を待たない.
- ✓ そこで本研究ではwebを用いたコミュニティツールの現状について整理・分類を行い，今後の発展に生かすこと，また新たな手法の提案，実装，評価を行う.

研究のフレームワーク

1.

Interview / Discussion / Broadcasting

- ✓ 合計4回ヒアリング調査(2009/07/10-11)
- ✓ 深夜ラジオ放送(2009/07/12)
- ✓ しまなみ風景遺産789「24時間ラヂオ」(2009/08/08-09)

Transcribing / Analysing / Modeling

2.

- ✓ 発話データのテキスト化
- ✓ グラフ分析を用いたテキスト内容の話題構造化
- ✓ バネモデルを用いた地域知ネットワークの表現

Feed Back

3.

- ✓ 評価
- ✓ アンケート
- ✓ 今後の課題

論文のレビュー

戸田浩之, 北川博之, 藤村考, 片岡良治, 2007: グラフ分析を利用した文書集合からの話題構造マイニング, 電子情報通信学会論文誌 D Vol. J90-D No.2 pp. 292-310,

背景・目的

- ✓ アクセス可能な情報の増大に伴う文書集合に対するユーザの要求が高まっている
  - 文書集合中の主要な話題は何かを理解したい
  - 特定の話題に関連する情報にアクセスしたい
- ✓ 文書へのアクセスの仕方
- ✓ 課題
  - 「話題へのアクセス時の課題」
    文書集合中に多くの話題が存在する場合，特定の話題へのアクセスや話題間のつながりを把握することが困難
  - 「文書へのアクセス時の課題」
    個々の話題（クラスタ）が多くの文書で構成されている場合，所望の文書にアクセスすることが困難.

関連研究

- ✓ PageRankアルゴリズムを用いた重要文抽出とキーワード抽出[Mihalcea,2004]
- ✓ グラフの中心性を用いたテキストの要約方法[Erkan,]
- ✓ WWWのコミュニティを抽出する手法
- ✓ ニュースストリームや長いテキストデータから内容が一貫している区間をサブトピックとして切り出す
- ✓ 時系列のニュースストリームを対象に，トピックの検出とトピックに関連するリソースの追跡
- ✓ 単に内容が類似しているだけでなく，より詳細な内容や別の視点から述べている文書を検索する手法

ノードの分類

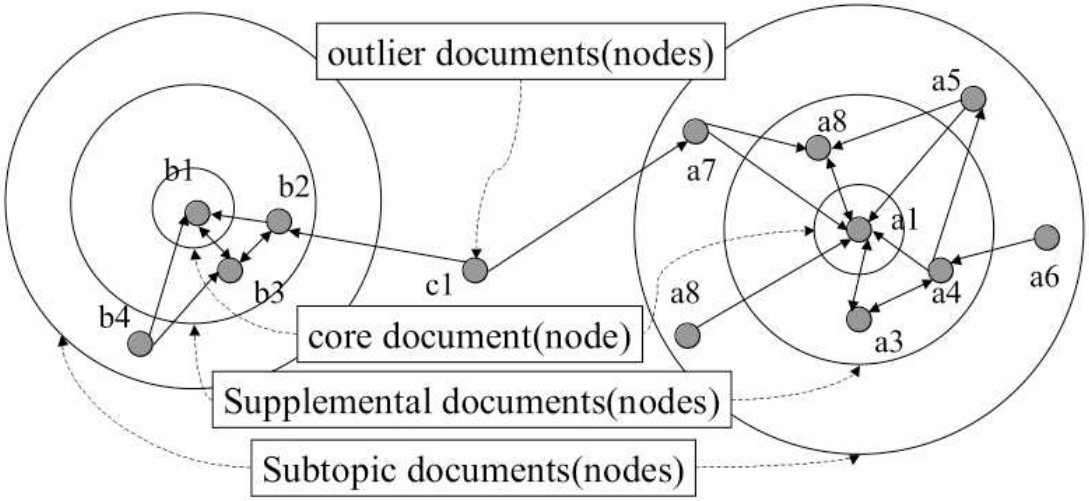


図 1　文書集合グラフの概念図

Fig. 1　Concept of document set graph structure.

✓ コアノード

- エッジでつながるどの隣接ノードよりも高い中心性をもつノード
- 隣接ノード群が示す話題の中心的な内容を最もよく示すと考えられる.

✓ サプリメンタルノード

- コアノードと強いつながりを持つノード
- コアノードの内容を補足する役割を持つ

✓ サブトピックノード

- コアノードもしくはサプリメンタルノードとつながりを持つノード
- 話題の中心と関連性はあるが，他文書とは異なる情報を示す文書.

✓ アウトライヤーノード

- 特定のノード群とはつながりを持たないノード.
- 他の文書とは内容が重ならない.

提案手法 - 文書集合グラフ構造構築

“Interested Reader Model”[Kamvar,2003] に基づくグラフ構造N

$$N = \frac{(A + d_{max}E - D)}{d_{max}}$$

■類似度上位p番目までのリンクを残す.

$$A_{i,j} = \begin{cases} sim(i,j) & \text{if } j \in TopSim_p(i) \\ 0 & \text{otherwise} \end{cases}$$

■類似度上位p番目までのリンクを残す.

$$N'_{i,j} = \begin{cases} N_{i,j}/l_{i,q} & \text{if } j \in TopLink_q(i) \\ 0 & \text{otherwise} \end{cases}$$

log tf-idf重み : 単語の出現頻度 × log{(単語を含む文書の頻度)$^{-1}$}

$TopSim_p(i)$: 文書iとの類似度が高い文書p件に含まれる文書集合

sim(i,j) : 文書をlog tf-idf重みによる単語ベクトルとして表現した場合の
文書i,jのコサイン類似度

✓ 類似度が高い文書間に対してのみエッジを設定

✓ すべてのノードから同じ本数(p本)のリンクが出ることになってしまう

提案手法 - 中心性スコアの算出法

PageRankを用いて，ノードの中心性スコアを与える.

$$S(V_i) = (1-d) \times \sum_{V_j \in IN(V_i)} (\frac{1}{|OUT(V_j)|} \times S(V_j)) + d$$

IN(V_i)：ノードV_iに対してリンクを張っているノードの集合.
OUT(V_i)：ノードV_iからリンクが張られているノードの集合.
d:ランダムジャンプ確率を設定するダンピングファクタ.

✓「話題の中心的な内容を含む文書は複数の文書それぞれと一定以上の類似度を持っている．」という仮定からPageRankを使用.

提案手法 - 話題構造マイニング

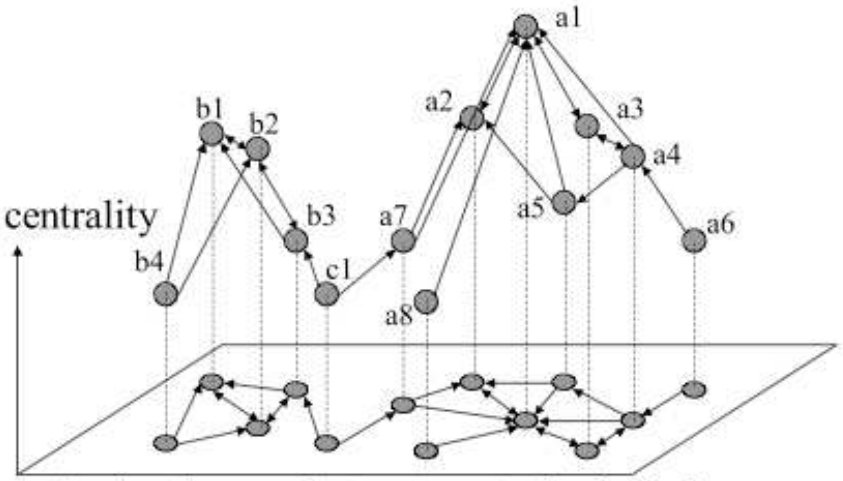


図 2　グラフ構造とノードの中心性を利用した文書集合の構造化

Fig. 2　Document set structure using graph structure and centrality score of each node in the graph.

- ✓ バネモデルを用いて，XY平面にグラフを配置
- ✓ Z軸上にノードの中心性スコアを割り当てる
- ✓ 山は異なる話題に対応する
- ✓ グラフ構造内でノードの4分類を適用

データの仕様

表 2　評価に利用した新聞記事テストセット及び主要話題リストの仕様
Table 2　Specification of the newspaper test set and main topic list for evaluation.

Name of set	Search query	# of docs.	# of labels	# of labeled docs.
murder	殺人	200	26	98
scandal	汚職 or 贈賄 or 収賄	200	22	170
kidnapping	誘拐	200	33	113
terrorism	テロ or 爆破 or 爆弾	200	28	105
s+t(scandal+terrorism)	—	400	50	274
s+k(scandal+kidnapping)	—	400	55	282
m+s(murder+scandal)	—	400	48	267
m+t(murder+terrorism)	—	400	54	203
m+k(murder+kidnapping)	—	392	56	205
k+t(kidnapping+terrorism)	—	399	61	219

表 3　実験条件
Table 3　Experimental conditions.

Parameter	Values
p: # of out-link from 1 node	2, 3, 4, 5, 6, 7
q: threshold for eliminating noisy edge(s)	0.5, 0.6, 0.7, 0.8, 0.9, 1

話題抽出精度の指標 F値

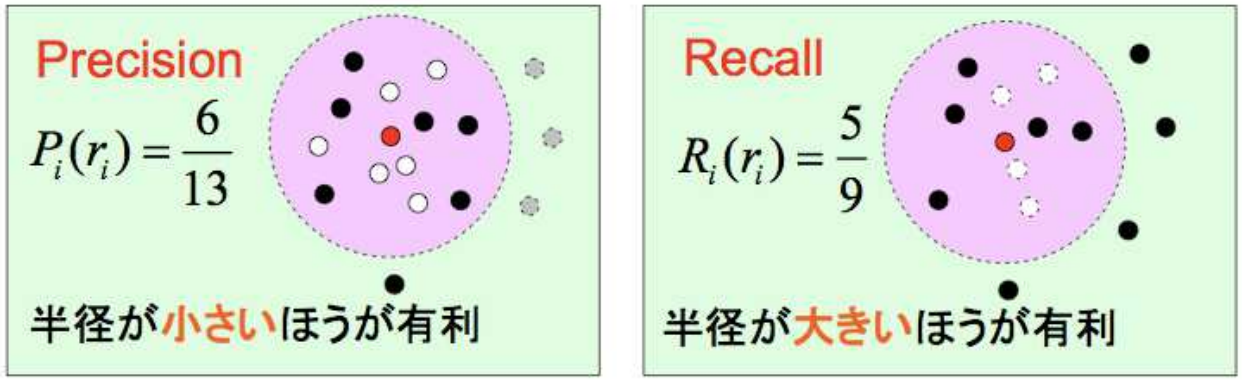
F値に基づく評価

全ノードのF 値の平均で埋め込みの良さを評価

F 値: Precision と Recall の調和平均

$$F_i(r_i) = 1/\{\alpha \frac{1}{P_i(r_i)} + (1-\alpha)\frac{1}{R_i(r_i)}\} \Rightarrow F = \sum_{i=1}^{N} \frac{F_i(\hat{r}_i)}{N}$$

通常 $\alpha = 1/2$



- ✓ Precision (適合率) とは、全検索結果に対しての、検索要求 (information need) を満たす検索結果の割合
- ✓ Recall (再現率) とは、検索要求を満たす全ドキュメントに対しての、検索要求を満たす検索結果の割合

評価

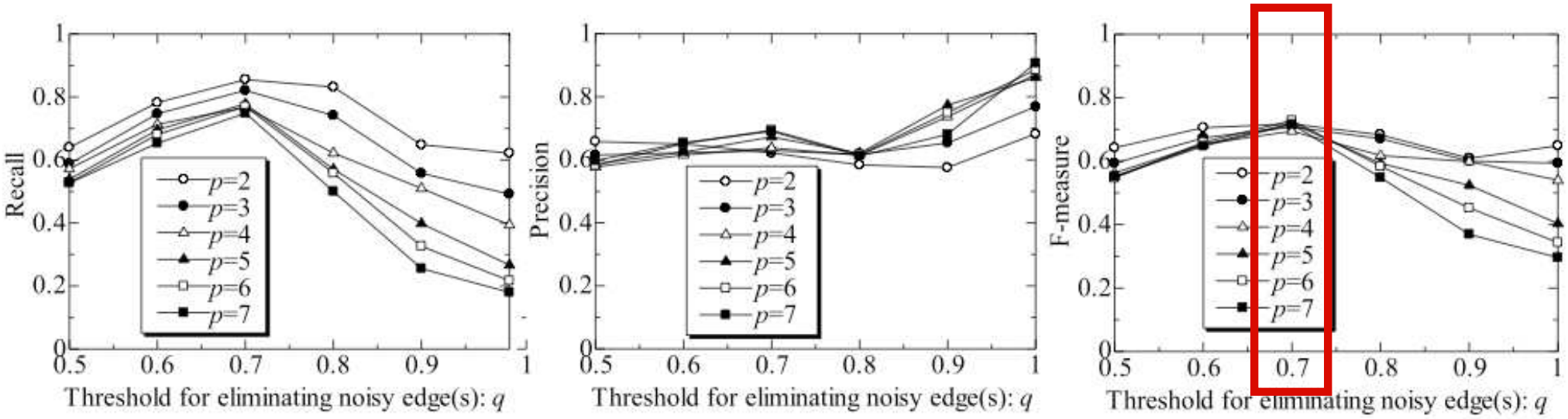


図 4　エッジ除去のしきい値（q）と話題抽出精度の関係

Fig. 4　Relationship between threshold for eliminating noisy edges (q) and accuracy of topic extraction.

q=0.7前後で精度のピークを
迎える結果となった.

✓エッジ除去のしきい値を変化させた場合の，話題抽出精度をみる

まとめ

- ✓ 話題構造マイニングの提案
- ✓ 文頭にあげた課題の解決
  - 文書集合中の主要な話題の提示
  - 文書間のつながりを視覚的に提示
  - 「話題の中心をよく示す文書」や「ノベルティの高い話題」等，文書を主題との関連性付きで提示
- ✓ 話題抽出精度について，ベースラインを上回る精度の提示
- ✓ クラスタリングについては既往手法に対して2つの指標に関しては値を上回る結果
- ✓ 文書集合の可視化結果の提示

研究に向けて

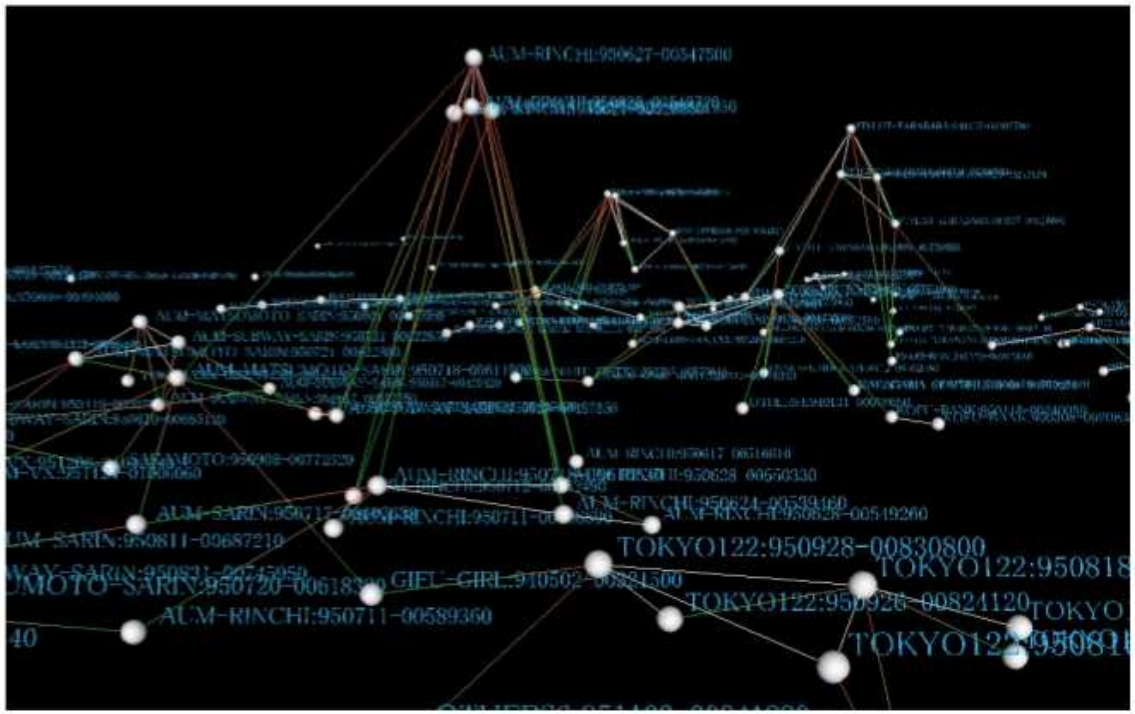


図 10　“murder” コーパスの可視化結果（$p = 3,\ q = 0.7$）
Fig. 10　Visualization example of corpus “murder” ($p = 3,\ q = 0.7$)

- ✓ 個人内に蓄積された発話内容を文書集合に見立て，検索を行いながら発言内容を行っていると仮定する.
- ✓ トピックの抽出可能性を算出出来るのではないか？
- ✓ 意識空間の可視化

既往研究 - 話題構造マイニング

- ✓ 戸田浩之, 北川博之, 藤村考, 片岡良治, 2007: グラフ分析を利用した文書集合からの話題構造マイニング, 電子情報通信学会論文誌 D Vol. J90-D No.2 pp. 292-310,
- ✓ 戸田浩之, 北川博之, 藤村考, 片岡良治, 2007: 時間的近さを考慮した話題構造マイニング, 電子情報通信学会 第18回データ工学ワークショップ論文集
- ✓ SD. Kamvar., D. Klein., CD. Manning., 2003: Spectral learning, International Joint Conference On Artificial Intelligence
- ✓ Blei, D., Ng, A., and Jordan, M., 2003: Latent dirichlet allocation
- ✓ Blei, D., and Lafferty, J.D., 2006: Dynamic Topic Models,Appearing in Proceedings of the 23 rd International Conference on Machine Learning
- ✓ 加藤義清, 赤石美奈, 堀浩一, 2009: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, The 23rd Annual Conference of the Japanese Society for Artificial Intel ligence,

参考文献 - 書籍

- ✓ Paul Thompson, 酒井 順子(訳), 2002/06:「記憶から歴史へ―オーラル・ヒストリーの世界」,青木書店
- ✓ 後藤 春彦, 田口 太郎, 佐久間 康富, 2005:「まちづくりオーラル・ヒストリー―「役に立つ過去」を活かし、「懐かしい未来」を描く」, 文化とまちづくり叢書
- ✓ 北研二, 確率的言語モデル