時間的近さを考慮した話題構造マイニング

戸田浩之[†,††]　北川博之[††,†††]藤村考[†]　片岡良治[†]

† 日本電信電話株式会社 NTT サイバーソリューション研究所
†† 筑波大学 システム情報工学研究科
††† 筑波大学 計算科学研究センター
E-mail: †toda.hiroyuki@lab.ntt.co.jp

あらまし　本稿では，発行日等のタイムスタンプを持つ文書の集合に対する話題構造マイニング手法について示す．話題構造マイニングとは，文書集合のマイニングにおいて，文書集合中の文書間類似度に基づいて構築したグラフ構造を考えることで，文書集合中の主要な話題やそれに該当する文書クラスタを抽出し，さらに複数の話題間の関係や特定の話題の中心的な内容に対する文書の位置付け (“中心的な内容を含む文書” や “関連情報を主に含む文書” 等) を特定する手法である．今回はこのグラフ構造を構築する際に，文書内容の類似度に加えて，時間的な近さを考慮する手法を提案する．また，この提案手法が話題抽出精度やクラスタリング精度に及ぼす影響および，抽出される話題構造の意味の変化について示す．

キーワード　テキストマイニング，話題構造マイニング，グラフ構造，PageRank

Topic Structure Mining using Temporal Co-occurrance

Hiroyuki TODA[†,††], Hiroyuki KITAGAWA[††,†††], Ko FUJIMURA[†], and Ryoji KATAOKA[†]

† NTT Cyber Solutions Laboratories, NTT Corporation
†† Graduate School of Systems and Information Engineering, University of Tsukuba
††† Center for Computational Science, University of Tsukuba
E-mail: †toda.hiroyuki@lab.ntt.co.jp

Abstract　This paper proposes a topic structure mining method for document set with timestamp. The topic structure mining is a text mining method using the graph structure based on document pair similarities in the document set. This method yields not only topic extraction from documents and clustering of documents but also extracting the relationship between clusters and meaning of each document in the cluster. Our method combines the temporal co-occurrance and document similarity to construct the graph structure. We also report the evaluation result and effect of this proposed method.

Key words　text mining, topic structure mining, graph structure, PageRank

1. ま え が き

近年アクセス可能な情報は増大し，多くのユーザが RSS リーダに蓄積される最新ニュースや，検索エンジンの結果等の “文書集合” に日々直面している．

この場合のユーザの要求としては，以下の 2 点が考えられる．

- 文書集合の中の主要な話題が知りたい
- 文書集合中の特定の話題に関連する情報にアクセスしたい

これらのユーザの要求を満たす手段として，重要キーワードの抽出により，文書集合中の話題を提示する手法 [13] [12] [9] や，クラスタリングにより，文書を分類提示する手法 [3] [7] [4] が挙げられる．

しかし，上記の手法で話題抽出や文書のクラスタリングができたとしても，我々は以下の問題点が存在すると考えている．

問題 1　話題へのアクセス時の問題
文書集合中に多くの話題 (クラスタ) 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事が困難

問題 2　文書へのアクセス時の問題
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事が困難

この問題に対して，我々は文書集合に対するマイニング手法として，話題構造マイニングを提案している [10]．この手法によると，上記の問題 1 については，文書集合中の主要な話題の

提示により話題の閲覧を支援したり，文書集合の関係可視化により文書や話題のつながりを容易に閲覧可能とし，話題間のつながりの強さ，つながりの内容の発見を支援する．また，問題 2 に対しても，主要な話題に関する文書クラスタを生成するだけでなく，クラスタメンバとなる個々の文書に，“話題の中心を最も良く示す文書” や “話題の中心とは関連するが，別の文書には含まれないノベルティの高い情報を含む文書” といったタイプ付けを行い，文書の選択的な閲覧を可能としている．

この手法では，文書間の関係を内容の類似度のみを用いて定義しているが，実際にユーザが扱う文書としては，ニュース記事やブログ記事等，発行日時が明記された文書が増えている．この場合，上記と同様にユーザは，文書集合中の主要な話題や，特定の話題にアクセスするという要求を持つと考えられるが，内容の類似度に加えて，時間的な近さを考慮することが重要になると考えられる．

これは，関連する話題の記事は近い時期に発行されることが多い為であり，時間的な近さを考慮することで，特定の話題に関する文書を精度良く集められると考えられる．例えば,「湾岸戦争」の記事と「イラク戦争」の記事は内容的には類似しているが，時間を考慮すればその差は明確である．また，時間情報を元に話題間のつながりを表現するという事も考えられる．

そこで，本稿では，発行日時を持つ文書集合中の所望の情報へ効率的にアクセスする手法として，話題構造マイニングにおいて，時間的近さを考慮する方法について示す．

基本的には，話題構造マイニングのグラフ構造を構築する段階で，文書内容の類似度に加えて，時間的な類似度 (以下，時間類似度) を考慮する事で実現する．これにより，文書間の内容の類似度が比較的高い場合でも，時間的に離れている場合には同じ話題について述べている可能性は低いと考え，類似度を低く見積もる．

以下，2 章では関連研究を示す．3 章では話題構造マイニングの概要について示し，4 章で提案手法について示す．5 章では評価について示し，6 章でまとめる．

また，本稿で言う話題とは，実世界で起こる一過性の高いイベントもしくはアクティビティを示しており，特別な説明がない限り，与えられた文書集合中の複数の文書で表現されたイベントやアクティビティの事を話題と呼ぶ．

2. 関 連 研 究

2.1 時間的近さを考慮したテキストマイニング

時系列を意識し文書集合の話題を特定しようとする取り組みとして石川らの研究 [15] が挙げられる．石川らは，ニュース記事のクラスタリングにおいて，最新記事群に含まれる話題を特定するために、忘却の概念を用いたクラスタリングを提案している。提案手法では、忘却の概念を導入することで、古い記事は新しく得られるどの文書とも類似性が低く、新しい記事同士がより高い類似性を持つというモデルを構築し、最新記事中の話題を特定している。重要と思われる最新の話題のみを選択的にクラスタリングする点で提案手法と類似している．また，Yang ら [11] は，ニュース記事群からイベント抽出を行う場合に，時間情報を導入することで，イベントの分類性が向上する事を報告している．

これらの手法はいずれもクラスタリングであるが，我々の手法は，クラスタリングだけでなくクラスタ間の関係や、クラスタ中の文書間の関係を抽出する点で異なる．

また，Cui ら [2] は，キーワードで与えられたトピックの活性度の時間変化を，文書ストリームとの内容類似度および時間的近さの両面から測る手法を提案している．この手法は文書集合の構造化とは異なるが，文書集合との内容類似度および時間的近さの両方を用いて話題の分析を行うという点で関連する．

2.2 グラフ構造を利用したテキストマイニング

近年，文書やセンテンス等の言語的な要素間の関係をグラフ構造で表現し，そのグラフ構造中のノードの中心性を利用することで，要素のランキングや上位の要素を抽出する手法が提案されている．

Mihalcea ら [8] は，要素間の類似度を元に構成されたグラフ構造にしばしば見られる特徴である「エッジ重みあり」,「エッジ方向性なし」の場合にも PageRank [1] が有益に働くことを，重要文抽出とキーワード抽出へ適用した実験から示している．また，Kurland ら [6] は，リンクのない文書集合において，言語モデルを用いたグラフ構造を構築する方法およびそれを利用した検索結果の再ランキングについて示している．

我々の手法も文書間の類似度を元にグラフ構造を構築しているが，今回の提案では文書間類似度に加え時間的な近さを考慮している点がまず異なる．また．上記の関連手法がノードの中心性スコアのみを用い，ノードのランキングや上位にランキングされたノードの抽出を行っているのに対して，我々の手法は中心性スコアだけでなく，グラフ構造を利用する事で，文書集合中に存在する話題を特定し，複数の話題間の関係や特定の話題に関連する文書が話題の中心に対してどのような位置付けであるかを明らかにする事を可能としている．

3. 話題構造マイニング

我々の提案する話題構造マイニングは，以下の 3 ステップで構成される．

（1） 文書集合を表現するグラフ構造を構築
（2） 構築したグラフ構造の個々のノードの中心性を算出
（3） グラフ構造と，ノードの中心性から話題構造を抽出

以下では，本稿での内容に関連性の高い，最初のグラフ構造の構築および，最後の話題構造の抽出について示す．

3.1 グラフ構造の構築

最初のステップのグラフ構造の構築では，文書集合から以下に示すノードとエッジ(注1)で構成される文書集合グラフ (以下，グラフもしくはグラフ構造と記す) を構築する．

ノード　文書集合中の各文書(注2)
エッジ　文書間の関係

(注1)：本稿では，エッジとリンクはほぼ同等の意味で用いる．ただし，方向性に関する議論を行う場合にはリンクという言葉を用いる．
(注2)：本稿では，1 つの文書は 1 つのノードと 1 対 1 で対応する．特にグラフ構造の説明を行う場合にはノードと言う表現を利用する．

このグラフ構造の構築には，Kamvar [5] らによって提案された “Interested Reader Model” をベースとして利用する．これは PageRank [1] の “Random Surfer Model” に類似したモデルであり，文書集合内の文書を次々に読み進める Reader を仮定している．このモデル中の Reader は以下のルールに従う．

- Reader が次の文書に遷移する場合にどの文書を選択するかは，今読んでいる文書に強く影響される．
- 現在の文書と類似した文書が存在しない場合，現在の文書にしばらく滞在する．

この Reader の遷移はマルコフ連鎖として表現され，上記の仮定に基づくグラフ構造は，以下の N で表現される．

$$N = (A + d_{max}E - D)/d_{max} \quad (1)$$

ここで，A はノード間の類似度を表現する隣接行列であり，E は単位行列を示す．D は対角行列であり，次の式で計算される．$D_{ii} = \sum_j A_{ij}$, また，d_{max} は D の要素のうち最大の値を取るものである．この式の最初の項 (A) が 1 つ目のルールを表現し，2 つ目の項 $(d_{max}E - D)$ で後者のルールを自己遷移の項として表現している．

ただし，ここで行列 A の全ての類似度を利用した場合，小さい類似度のみの関係がノイズとなる事が考えられる．そこで，有益な関係のみを選択的に抽出する為に，以下の 2 段階の操作を行う．

（1） それぞれのノードからのアウトリンク数を限定する．
それぞれのノードにおいて，最も類似度の高い p 個のノードへのみリンクを設定する．

（2） 極小さい重みで設定されたリンクを除去する．
それぞれのノードにおいて，自己遷移を含むアウトリンクをその重みの降順に並べ，リンク除去係数 q を越えるまでそれぞれのリンクの重みを加算し，q を越えた時点でそれまでに加算対象となったリンクのみを設定する．

最初の段階で，規定数のアウトリンクを設定することで，一部のノードにごく小さい類似度を持つアウトリンクが存在してしまう問題を回避するために，2 つ目の段階を設けている．

3.2 話題構造の抽出

最後のステップでは，最初のステップで得られたグラフ構造を XY 平面に配置し，2 番目のステップで得られたノードの中心性 (提案手法では，PageRank を利用して算出) を Z 軸上に割り当てた 3 次元のグラフ構造を考える．例を図 1 を示す．ここで，ノード “ax”(“bx”) は，話題 “a”(“b”) に関するノードである事を示す．

中心性の定義によれば，多くのエッジが存在するエリアのノードは高い中心性を持つ．このようなエリアはまた，そのエリア内での状態遷移確率が高く，ノード間の関連性も高い．つまり，そのようなエリアは同じ話題に関連するノードで構成される．したがって，図 1 のそれぞれの山は，それぞれ個々の話題に対応すると考えられる．

我々はこの構造を利用して，文書集合からの話題抽出，クラスタリングを行う事を提案している．つまり，図 1 のそれぞれの山の頂点となる文書を集めることで，文書集合中の主要な話題を抽出し，また，そこで抽出した文書を中心とした山状の構造を構成する文書を集めることでクラスタを生成している(注3)．

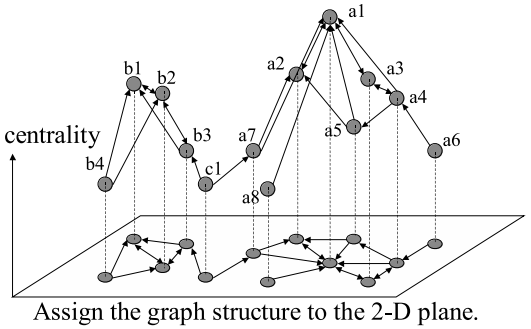

図 1 グラフ構造とノードの中心性を利用した文書集合の構造化
Fig. 1 Document set structure using graph structure and centrality score of each node in the graph

ただし，前節で示したアウトリンク数 p やリンク除去係数 q の設定，さらに今回導入する時間類似度の設定によっては，無駄なリンクが多くなり山状の構造が隠れ，話題抽出が出来なかったり，逆にリンクがスパースになり過ぎて必要十分な構造が得られない場合もある．

4 章では，上記に示す手法によって行った話題抽出およびクラスタリングの精度が時間類似度を導入することでどのように変化するかを評価する．

また，この山に含まれるノードの位置に応じて，それぞれの文書は以下の 4 タイプに分類できる．これを利用することで，文書やクラスタへの効率的なアクセスを支援する．

コアノード (文書) エッジでつながるどの隣接ノードよりも高い中心性を持つノード．隣接ノード群が示す話題の中心的な内容を最も良く示す文書 (a1,b1).

サプリメンタルノード (文書) コアノードと強いつながりを持つノード．コアノードが示す話題の中心を補足する文書 (a2,a3,a4,b2,b3).

サブトピックノード (文書) 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 (a5,a6,a7,a8,b4).

アウトライヤーノード (文書) 特定のノード群とつながりを持たないノード．他の文書とは内容の重ならない文書 (c1).

例えば，クラスタの中心的な内容を知りたければ，コアノードで表現される文書を読むことで概要が理解できたり，サプリメンタルノードやサブトピックノードを共有する山が存在すれば，それらの山は相互に関連する話題について述べられていることがわかり，また，その関連性は共有された文書を参照する事で理解することができる．

4. 提 案 手 法

本章では，時間類似度の定義およびそれを利用した文書間の類似度算出法について示す．ここで示す類似度を，グラフ構造

(注3)：実際には，一つの頂点のノード (コアノード) と直接もしくは間接的なリンクで接続されるサプリメンタルノード，サブトピックノードを集め，一つのクラスタとして扱う．また，アウトライヤーノードはいずれのクラスタにも属さない文書と見なす．

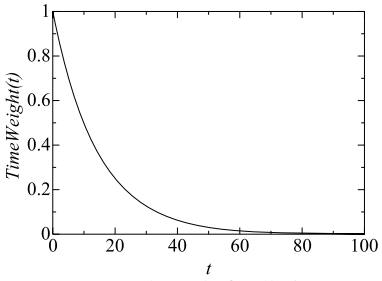

図 2　時間類似度の推移

Fig. 2　Change of Time Co-occurrance Similarity

構築の際に利用することで，時間的近さを考慮した話題構造マイニングを可能とする．

4.1　時間類似度の定義

ここでは,「文書間のタイムスタンプが一定の時間離れる毎に,一定の割合で類似度が減少する」との仮定に基づき時間類似度を定義する．この仮定は以下の式で表現できる．

$$\frac{d}{dt}TimeWeight(t) = -\lambda \times TimeWeight(t) \qquad (2)$$

ここで，λ はタイムスタンプの差の拡大による類似度の逓減の程度を示す定数，t は二つの文書のタイムスタンプの差を示し，$TimeWeight(t)$ は，二つの文書のタイムスタンプの差が t だった場合の時間類似度を示す関数である．

この常微分方程式を解くと，タイムスタンプの差が t の場合の時間類似度 $TimeWeight(t)$ は以下の式で表現できる．

$$TimeWeight(t) = T_0 \times \exp(-\lambda t) \qquad (3)$$

ここで，T_0 は，タイムスタンプの差が 0 の場合の重みであり，タイムスタンプの差 t が大きくなるにつれ，時間類似度が減少し，最後には 0 に限りなく近づく．

ただ，このままの式では，類似度逓減の度合を直感的に設定することが困難である為，以下の様に式を変形し逓減の度合を示すパラメータを $t_{1/2}$ とする．

$$TimeWeight(t) = T_0 \times \exp(-\frac{0.693}{t_{1/2}}t) \qquad (4)$$

上式において，$t_{1/2}$ は，時間類似度が 50%になるタイムスタンプの差 (半減期) を示している．図 2 に，$T_0 = 1, t_{1/2} = 10$ とした場合の時間類似度の推移を示す．

4.2　時間類似度を考慮した類似度の定義

以下に文書内容に基づく類似度と，時間類似度を考慮した文書間類似度 $sim(i,j)$ の定義を示す．

$$sim(i,j) = sim'(i,j) \times ((1-\alpha) + \alpha \times TimeWeight(t(i,j))) \qquad (5)$$

ここで，$sim'(i,j)$ は文書 i と j の文書内容に基づく類似度を示し，α は時間類似度の重みを調整するパラメータである．

図 3 に，この式によって時間類似度による重みがどのように適用されるかを示す．

また，類似度 $sim'(i,j)$ は，文書のキーワードベクトル間のコサイン類似度で算出する．それぞれのキーワードベクトルの

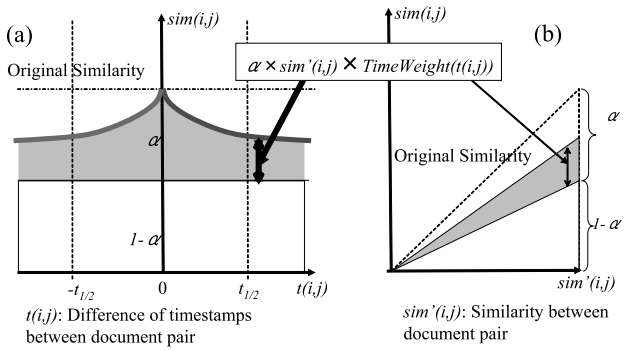

図 3　類似度の変化: (a) タイムスタンプの差が変化した場合，(b) 内容類似度が異なる場合

Fig. 3　Changes of similarity: (a) the case: the difference of timestamps is varied, (b) the case: original similarity is varied.

要素には，文書を形態素解析器 ChaSen(注4) で解析し，名詞もしくは未知語と判定された語を利用した(注5)．各要素の重みは，以下に示す logarithmic tf-idf で算出した．

$$w_{x,j} = \log(1 + tf_{x,j}) \times \log(N/df_x) \qquad (6)$$

$w_{x,j}$ は文書 j 中の単語 x の logarithmic tf-idf 重みを示し，$tf_{x,j}$ は文書 j 中での単語 x の出現頻度，df_x はコーパス全体での語 x の出現頻度，N は，コーパス全体の文書数を示す．

従来の話題構造マイニングのグラフ構築プロセスでは，文書間類似度として，内容に基づく類似度のみを利用していたが，今回の提案手法では，グラフ構造の元となる文書間類似度の計算に本節で定義した類似度 $sim(i,j)$ を利用する事で，時間類似度を考慮した文書集合グラフの構築を行い，文書の時間的近さを考慮した話題構造マイニングを行う．

5.　評　　価

本評価では，時間類似度を考慮することによる「話題抽出，クラスタリングの精度」および「抽出された話題構造の性質」の変化について明らかにし，その原因を考察する．

5.1　評価リソース

今回の評価では，新聞記事コーパスに対して検索を行い得られた検索結果で評価を行った．利用した新聞記事は 1994 年および 1995 年の毎日新聞の記事であり，約 20 万件の新聞記事から構成されている．このコレクションを全文検索システムに登録し，キーワード検索で得た結果それぞれ上位 200 件を 1 つのコーパスとする．ここでは，4 つのコーパスを作成した．実際の評価用のテストセットとしては，この 4 つのコーパスをそれぞれ単独で利用した 4 セットと，それぞれ 2 つのコーパスを組み合わせた 6 セットの合計 10 セットを利用した．テストセットの仕様を表 1 に示す．組み合わせたテストセットのうち，m+k および k+t は，元となったテストセットに重複した文書が存在した為，文書数が 400 より少ない．今回は一過性の高いイベントに注目した為，それらのイベントが起こりやすいと考えられ

(注4)：http://chasen.naist.jp/hiki/ChaSen/

(注5)：ただし，未知語の連続は一つの語として扱った．

表 1　評価に利用した新聞記事テストセットおよび主要話題リストの仕様
Table 1　Specification of the newspaper test set and main topic list for evaluation

Name of set	Search query	# of docs.	# of labels	# of clustered 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

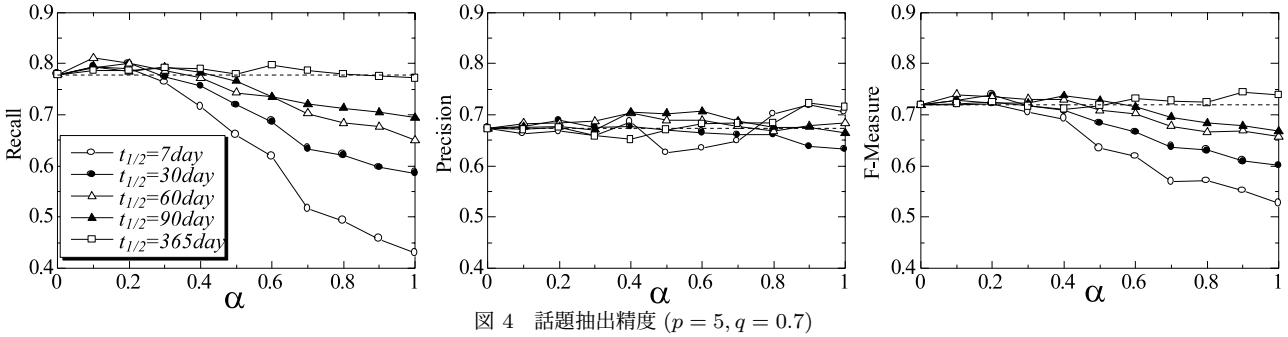

図 4　話題抽出精度 ($p = 5, q = 0.7$)
Fig. 4　Accuracy of Topic Identification($p = 5, q = 0.7$)

る分野の言葉をテストセット作成用の検索キーワードとした．

コーパスの各文書に対して本稿の第 1 著者がラベル付けを行った．各々のラベルは文書中で主に述べられている話題を示す．この結果を元に，各テストセット内で，2 文書以上で述べられている話題のみを集めた主要話題リストを作成した．Table1 中の “# of Labels” に，それぞれの主要話題リストに含まれる話題の数を示す．このリストを話題抽出の評価における正解データとして利用する．また，主要話題リストのそれぞれの話題に関連する文書を集めたデータをクラスタリング評価における正解データとして利用する．ここで，1 文書のみで述べられている話題は主要話題リストに含まず，またその話題に関連する文書もクラスタリング評価用の正解データには含まない．

精度の評価基準としては，話題抽出では適合率，再現率，F 値，クラスタリングでは F-Score [14] を利用した．F-Score はクラスタリングの精度測定指標の 1 つであり，それぞれの正解クラスタと最も類似したクラスタの適合率をクラスタサイズによる重み付きで平均した値である．

5.2　実 験 条 件

ここで，パラメータ p と q は，3.1 で示したように，グラフ構造構築に関するパラメータである．今回の評価では，時間類似度を考慮しない場合に，話題抽出およびクラスタリングの両方で高い精度を示した条件 “$p = 5, q = 0.7$” を元に条件の設定を行った．また，この条件で時間情報を考慮しない条件をベースラインとした．

また，時間類似度を変化させる為のパラメータとして，時間類似度の半減期を示す $t_{1/2}$ と時間類似度の強さ (時間重み) を決定する α がある．今回の実験では，新聞記事を利用しており，その発行日時は 1 日毎であるため，時間的な差を示す値の単位は日とする．また，T_0 は 1 とした．

以下に示す話題抽出およびクラスタリングは，時間類似度を考慮して構築したグラフ構造に対し，3.2 で示した手法を用いて行っている．

また，特に明記する場合を除き，それぞれの評価値は，10 個のテストセットを用いて得られた評価値の平均を示す．

5.3　話題抽出精度について

まずベースラインと同じ条件 ($p = 5, q = 0.7$) において，時間類似度を導入した場合の結果を図 4 に示す．グラフでは，縦軸にそれぞれの評価値，横軸に時間重みの強さを示す．本稿で示すグラフ内の破線はベースラインの評価値を示す．

話題抽出の適合率は，半減期および時間重みのいずれを変化させた場合にも大きな変化は見られない．一方，再現率は，時

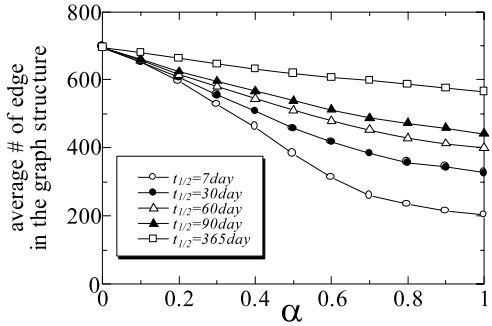

図 5　エッジ数の変化 ($p = 5, q = 0.7$)
Fig. 5　Change of number of links($p = 5, q = 0.7$)

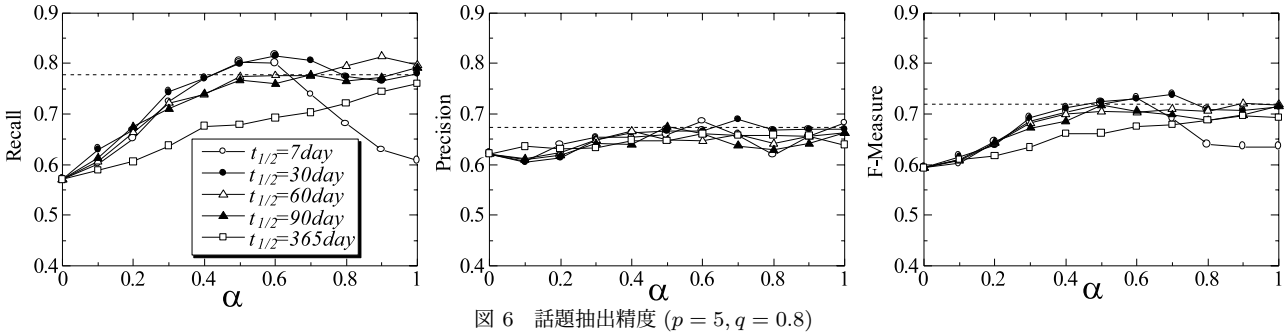

図 6　話題抽出精度 ($p = 5, q = 0.8$)

Fig. 6　Accuracy of Topic Identification($p = 5, q = 0.8$)

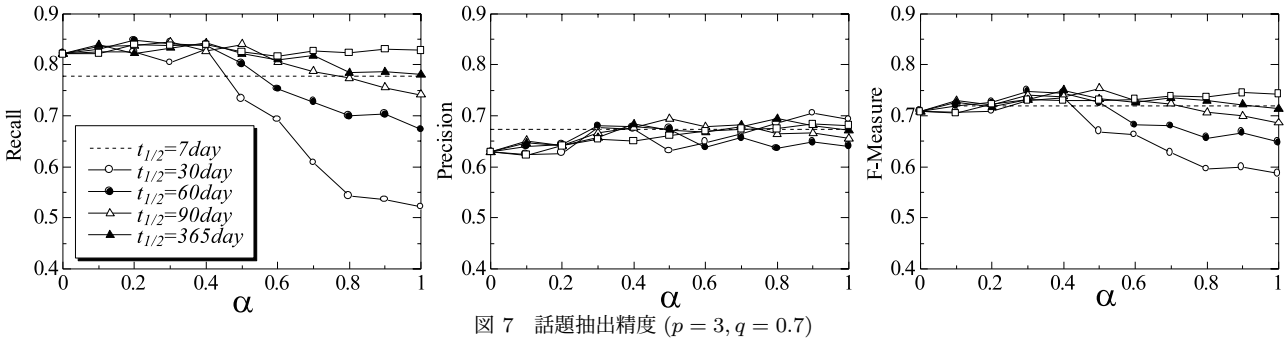

図 7　話題抽出精度 ($p = 3, q = 0.7$)

Fig. 7　Accuracy of Topic Identification($p = 3, q = 0.7$)

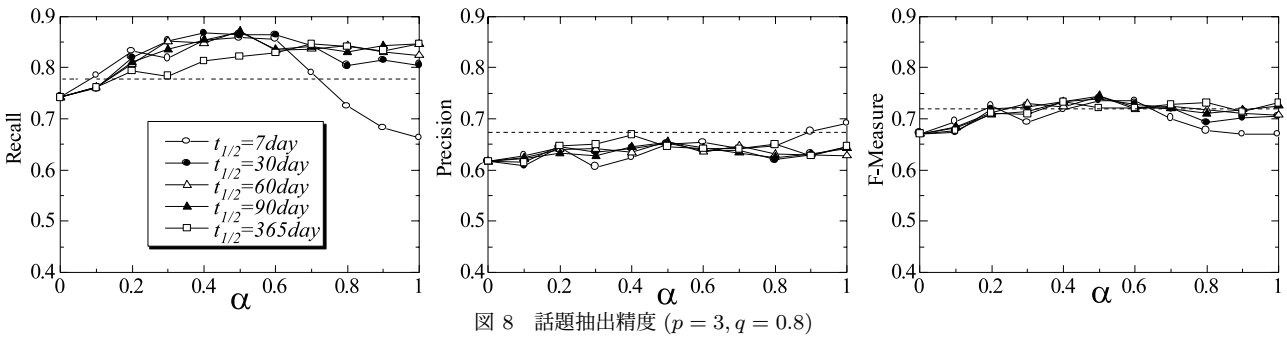

図 8　話題抽出精度 ($p = 3, q = 0.8$)

Fig. 8　Accuracy of Topic Identification($p = 3, q = 0.8$)

間重みが弱い場合には，変化が見られなかったが，時間重みを強くするにつれ，また半減期を短くするにつれ，低下している．

時間重みを強くする事および半減期を短くすることは，ともに時間類似度を強く導入する操作であり，それにより，話題抽出の精度が低下する事を示している．

この原因の究明するため，構築されたグラフ構造のエッジ数を調査した結果を図 5 に示す．ベースラインでのエッジ数はグラフの左端の点 (695) である．この結果から，時間重みを強く導入した場合，つまり，半減期を短くした場合，および時間重みを強くした場合に，リンク数が減少している事がわかる．

このリンク数の減少は話題構造マイニングのリンク除去プロセスが関係している．3. 1 で示したように，リンク除去プロセスでは，ノイズとなるリンクの影響を除去するために，小さい重みしか持たない文書間のリンクを除去している．今回，時間類似度を導入した事により，内容が類似していてもタイムスタンプに差がある場合に類似度が低下し，有益なリンク同士でも重みの差が大きくなっている．この為，時間類似度を導入していない条件 (p, q) をそのまま利用した場合，有益なリンクが除去され，再現率が低下したのではないかと考えられる．

この有益なリンクの除去を抑える一つの方法は，リンク除去係数 q の値を大きく設定し，ある程度小さい重みのリンクも残す事である．また，上記で述べたように，全体的な傾向として類似度の値の差が大きくなった場合，グラフ構造中の自己遷移 ($d_{max}E - D$) が高い割合を占めるようになり，自己遷移以外のリンクが除去される状況が考えられる．これを回避するには，1 ノード辺りのリンク数 p を小さくし d_{max} を小さくする事が考えられる．

以上を踏まえ，有益なリンクの除去を防ぐ為に，リンクの除去を少なくした条件 ($p = 5, q = 0.8$) および 1 ノード辺りのリンク数を小さくした条件 ($p = 3, q = 0.7$)，さらに両方の対処を行った条件 ($p = 3, q = 0.8$) で評価を行った．結果を図 6-8 に示す．

“$p = 5, q = 0.8$” の場合には，時間類似度を導入しない条件で再現率が低い．これは時間類似度を考慮しない場合，リンク除去が十分に行われておらず，ノイズとなるリンクが残存し，いくつかのコアノードが埋もれてしまったと考えられる．これは，グラフ構造中の総リンク数がベースラインの 1.5 倍 (1054) である事，抽出される平均話題数がベースラインより少ないこ

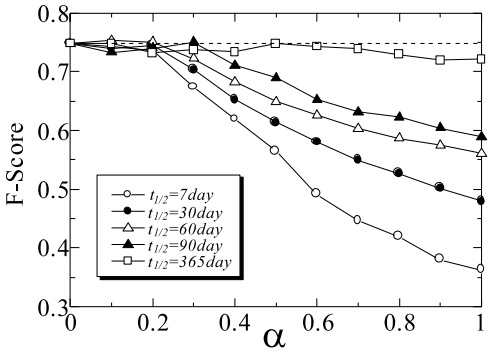

図 9　クラスタリング精度 ($p = 5, q = 0.7$)

Fig. 9　Accuracy of Clustering Result($p = 5, q = 0.7$)

と (ベースライン:44.4，正解データ:43.4，本条件:32.3) からも明らかである．しかし，時間類似度を導入することで，不要なリンクが除去され，これにともない再現率は向上し，時間重みが 0.5 辺りではベースラインの精度を上回った．ただし，F 値において，ベースラインとの間に有意な差はなかった．

“$p = 3, q = 0.7$” では，再現率は元々ベースイラインを上回り，時間類似度を導入しても同傾向である．一方，適合率は時間類似度を導入しない状態では，ベースラインを下回るが，時間類似度を導入する事で，ベースラインに近い値を示す．半減期を 30〜90 に設定し，時間重みを 0.5 程度にした場合には，F 値でベースラインを上回り，その差は有意傾向であった ($t_{1/2} = 60, \alpha = 0.5$ の場合，両側検定: $t(9) = 1.9270, .05 < p < .010$)．この条件で得られるグラフ構造中のリンク数は “$p = 5, q = 0.7$” の場合と同程度少ないが，有益なリンクが残存したため，話題抽出精度がある程度高くなったものと考えられる．これは，抽出される話題数が多い事にもあらわれている ($p = 3, q = 0.7$ の場合: 47.1，$p = 5, q = 0.7$ の場合: 37.4)．

一方，“$p = 3, q = 0.8$” の場合，再現率は，“$p = 3, q = 0.7$” の条件と同様に，時間類似度を導入しない場合にもある程度高い．この条件では，再現率は時間重みを 0.5 周辺とした場合，ベースラインと比較して最高 10%程度向上している．これらの条件では，抽出話題数がベースラインより 20%程度多い事も影響していると考えられる．一方，適合率は時間類似度を考慮していない場合をやや上回るものの，ベースラインよりは低い．しかし，$t_{1/2} = 90, \alpha = 0.5$ の場合には，話題抽出の F 値が 0.7443 程度を示し，t 検定の結果，ベースラインの結果との間には有意な差があった (両側検定: $t(9) = 2.5723, .01 < p < .05$)．

5.4　クラスタリング精度

図 9 にベースラインの条件に対して時間類似度を導入した場合のクラスタリングの評価結果を示す．話題抽出精度と同様に時間重みを増やす程，また半減期を短くする程，精度が低下している．このように話題抽出と同様の傾向になるのは，提案手法におけるクラスタリングが話題抽出で抽出するコアノードを元にしている為である．この為，図 5 で示したリンク数の減少が精度低下に影響していると考えられる．

実際この条件でクラスタリングを行った場合に，いずれかのクラスタのメンバとして集められた文書数 (アウトライヤと見なされた文書を除いた文書数) は，ベースラインと比較して，20%〜50%程度少ない (ベースライン:190，正解データ:193.6)．

そこで，話題抽出と同様に，リンクの除去を少なくした条件 ($p = 5, q = 0.8$) および 1 ノード辺りのリンク数を小さくした条件 ($p = 3, q = 0.7$)，さらに両方の対処を行った条件 ($p = 3, q = 0.8$) で評価した．結果を図 10 に示す．

このうち，“$p = 3, q = 0.7$” は，“$p = 5, q = 0.7$” の場合と同様に精度が低下している．この条件は，“$p = 5, q = 0.7$” と同様にグラフ構造中のリンク数が少ない条件である．リンクの質の違いで，クラスタメンバは “$p = 5, q = 0.7$” と比較してやや多く集められているが，ベースラインと比較すると最大で 40%程度少なく，これが精度低下の原因であると考えられる．

それ以外の 2 つの手法では，半減期が今回実験した範囲の中程 ($t_{1/2} = 30$〜90) で，かつ時間重み $\alpha = 0.5$ 前後でベースラインを上回る精度を示している．また，その中のいくつかの条件では，ベースラインとの精度差に有意な差があり ($p = 3, q = 0.8, t_{1/2} = 60, \alpha = 0.5$ の場合; 両側検定: $t(9) = 2.534, .01 < p < .05$)，時間類似度を利用しない場合と比較して高い精度を得ることがわかった．また，クラスタメンバの数もベースラインと同等以上であった．

5.5　精度評価のまとめ

以上の評価結果より，今回利用した新聞記事を用いた評価では，半減期 $t_{1/2}$ を 30〜90 程度に設定し，かつ時間重み α を 0.5 前後に設定した場合，話題抽出およびクラスタリングの精度がベースラインを有意に上回る事がわかった．

ただし，時間情報を利用する場合には，時間を考慮しない場合と比較して，リンク数 (p) を少なくし，かつリンク除去係数 (q) を高めに設定する必要があった．これは，時間類似度を導入する事で，文書間の類似度の差が大きくなり，リンク除去プロセスで有益なリンクが除去される傾向がある為だと考えられる．

一方，半減期を極端に短く設定した場合や時間重みを極端に強く設定した場合には，精度が低下し，逆に半減期を長く設定した場合には，ベースラインからあまり変化しなかった．

また，提案手法によって大きく精度が向上したテストセットとして，“tero” が挙げられる．元々の精度は，他のテストセットの精度と比較して低い傾向にあった (ベースラインで，話題抽出の F 値:0.58, F-Score:0.65)．これは，テストセット中に存在する話題の多くが類似した話題 (パレスチナ問題) に関係し，文書内容に基づく類似度のみでは十分に話題の分類が出来なかった為である．これに対し，提案手法を利用することで，それぞれの指標で 0.1 ポイント以上の向上が見られ ($p = 3, q = 0.7, t_{1/2} = 60, \alpha = 0.5$ の場合，話題抽出の F 値:0.71, F-Score:0.75)，提案手法がうまく作用した例と言える．

5.6　時間類似度利用による話題構造の変化

本節では，時間類似度を利用した事で，抽出される情報がどのように変化したかについて示す．ここでは特にコアノードとして抽出される文書の変化について示す．

我々の提案する話題構造マイニングにおいて，コアノードとして抽出される文書 (コア文書) は，クラスタを代表する文書である．時間類似度を考慮せずに，新聞記事を対象にマイニングを行った場合，抽出されるコア文書は，クラスタ内の話題を

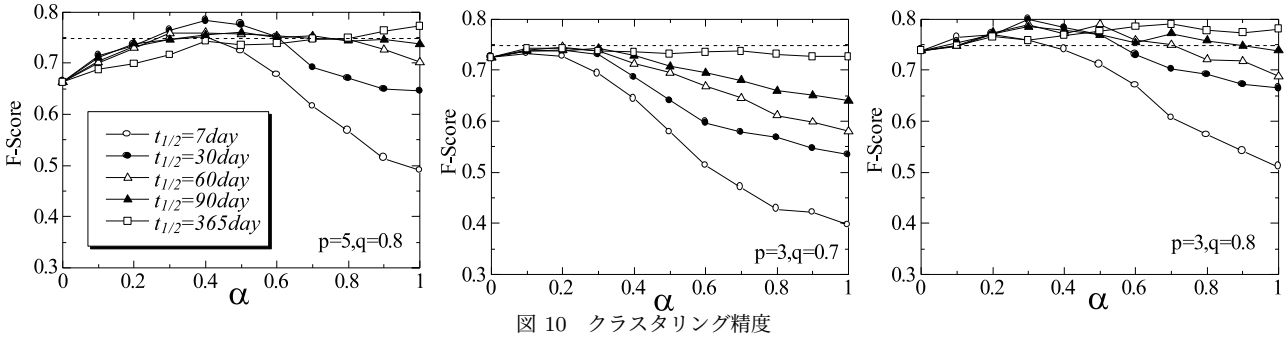

図 10　クラスタリング精度

Fig. 10　Accuracy of Clustering Result

網羅するような文書であったが，話題が起こっている最中の記事というよりは，話題として収束する段階の記事である事が多かった．具体的に言うと，事件に関するクラスタから抽出されるコア文書には裁判に関する文書が抽出される場合が多い．確かに裁判に関する記事では，事件の全体像や登場人物について詳しく記述されており，内容のみを考慮すると正しい文書である．ただし，そのコア文書が抽出されたクラスタを見ると，事件を伝える情報が多く存在し，裁判に関する文書は期間があいた後で発行されている事が多い．

一方，提案手法により，時間類似度を導入した手法で抽出されるコア文書は，ある話題に関連する文書 (記事) が次々と発行されている中で発行され，かつ多くの内容を包含する情報を抽出する事が出来ており，時間類似度を利用した効果が見られる．

具体例としては，今回の “osyoku” セット内に含まれる「つくば市の汚職」に関する話題が挙げられる．この話題に関しては，1994 年 11 月 2，3，4 日に汚職の発覚に関する記事が発行され，11 月 23 日に起訴，翌年 1995 年 1 月 20 日に初公判の記事と続いている．従来手法では，最後の初公判の記事がコアノードとして抽出されたのに対し，時間類似度を考慮した手法では，話題が盛り上がっている 11 月 3 日の記事を抽出している．どちらも事件の主要な人物が登場し，事件の概要について書かれている為，文書群の内容を知りたいという目的ではどちらの文書でも問題ないが，話題が盛り上がっている部分の記事を抽出している点が，これまでとの違いであり，提案手法の特徴的な点であると言える．

6. ま と め

本研究では、タイムスタンプを持つテキスト集合に対して話題構造マイニングを適用する手法を提案し，時間類似度を考慮する事が話題抽出およびクラスタリングの精度に及ぼす影響，およびその精度変化の要因について分析した．また，提案手法によって抽出される情報が，時間類似度を考慮しない場合と比較してどのように変化したかについて検証を行った．

以上より以下の結果を得ることが出来た．

- 適切なパラメータをセットすることで，話題抽出，クラスタリングの精度が向上する事
- 時間類似度を利用する場合には，グラフ構造を構築する段階で，時間を考慮しない場合と比較して，1 ノード辺りのアウトリンク数を少なくし，かつ不要なリンクの除去の度合を弱め (リンク除去係数を高め) に設定する必要がある事
- 時間類似度を利用した場合，コア文書として，関連する情報が短期間で連続的に発行されるような，話題が盛り上がっている時期の文書を選択する傾向にある事

今後は，今回注目した一過性の話題だけでなく，例えば，ライフスパンが長かったり，周期的に表れる話題等を考慮した手法の改良とその評価手法の検討，より大規模かつ多様なテストセットを利用した評価を行う予定である．

文　献

[1] S. Brin, and L. Page, “The anatomy of a large-scale hypertextual Web Search Engine,” Proc. of WWW7, pp.107–117, 1998.

[2] C. Cui, and H. Kitagawa, “Topic Activation Analysis for Document Streams Based on Document Arrival Rate and Relevance,” Proc. of SAC ’05, pp.1089–1095, 2005.

[3] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, “Scatter/Gather: a cluster-based approach to browsing large document collections,” Proc. of SIGIR ’92, pp.318–329, 1992.

[4] X. He, C. Ding, H. Zha, and H. D. Simon, “Automatic Topic Identification Using Webpage Clustering,” Proc. of ICDM ’01, pp.195–202, 2001.

[5] S. D. Kamvar, D. Klein, and C. D. Manning, “Spectral Learning.” Proc. of IJCAI ’03, pp.561–566, 2003.

[6] O. Kurland, and L. Lee, “PageRank without hyperlinks: Structural re-ranking using links induced by language models,” Proc. of SIGIR ’05, pp.306–313, 2005.

[7] A. Leuski, “Evaluating document clustering for interactive information retrieval,” Proc. of CIKM ’01, pp.33–40, 2001.

[8] R. Mihalcea, and P. Tarau, “TextRank: Bringing Order into Texts,” Proc. of EMNLP ’04, pp.404–411, 2004.

[9] H. Toda, and R. Kataoka, “A search result clustering method using informatively named entities,” Proc. of WIDM ’05, pp.81–86, 2005.

[10] H.Toda, K. Fujimura, R. Kataoka, and H. Kitagawa. “Topic Structure Mining using PageRank without Hyperlinks”, Proc. of ICADL ’06, pp.151-162, 2006.

[11] Y. Yang, T. Pierce, J. Carbonell, “A Study on Retrospective and On-Line Event Detection,” Proc. of SIGIR ’98, pp.28–36, 1998.

[12] H. Zeng, Q. He, C. Zheng, W. Ma, and J. Ma, “Learning to cluster web search results,” Proc. of SIGIR ’04, pp.210–217, 2004.

[13] O. Zamir, O. Etzioni, and A. Grouper, “Grouper: A Dynamic Clustering Interface to Web Search Results.” Proc. of WWW8, pp.1361–1374, 1999.

[14] Y. Zhao, and G. Karypis, “Evaluation of Hierarchical Clustering Algorithms for Document Datasets” Proc. of CIKM ’02, pp.515–524, 2002.

[15] 石川佳治，北川博之, “忘却の概念に基づくクラスタリング手法の改良方式,” 日本データベース学会 Letters Vol.2, No.3, 2003.