トピックに注目した
文書検索結果の構造化に関する研究

システム情報工学研究科

筑 波 大 学

2007 年 7 月

戸田 浩之

概要

近年，コンピュータネットワークの普及により，アクセス可能な情報は増大し，情報を取得する際に検索システムを利用する事が，多くなっている．検索の対象となる情報としては文書 (テキスト)，画像，映像，音声，音楽等様々な種類が存在するが，本論文ではこのうち文書情報の検索に関する内容を取り扱う．

現在，文書情報を検索するシステムとしてキーワードベースの検索システムが幅広く利用されている．代表的なものとして Google や Yahoo!等の Web 検索エンジンがあげられるが，イントラネット内での組織内文書の検索，PC 内の文書検索にも利用されており，コンピュータを利用する限り，なくてはならないものとなっている．

キーワードベースの検索システムを利用する場合の手順としては，まずユーザは適当なキーワードをシステムに入力し，そのキーワードを含む検索結果のリストを取得する．そして，そのリストの中から，個々の検索結果のタイトルやスニペットの情報，場合によっては本文を閲覧する事で，真に欲しい情報を選別し，取得する．

しかし，最後の検索結果のリストから真に欲しい情報を選別する作業は，しばしばユーザにとって困難かつ時間がかかる作業となる．そのような状況に陥る主なケースとして，以下に示す二つのケースが考えられる．一つ目のケースは，特定の文書を求めて検索を行うが，検索条件があいまいとなってしまう場合である．検索条件に指定されたキーワードが多義性を持っていたり，ユーザが明確な検索条件を作成できないという原因が考えられる．このような場合には，ユーザが求める情報は大量の検索結果の中に埋もれてしまい，その中からユーザは所望の情報を探さなければならない．また，二つ目のケースとしては，元々の目的が検索結果全体から，概要や特徴的な情報を抽出しようとする場合である．このような場合にも，ユーザは所望の情報を取得するために，膨大な検索結果を閲覧する必要がある．

本論文では，上記で指摘した問題への解決策として，検索結果中に含まれるトピックに注目した検索結果の構造化手法を提案する．

上記一つ目のケースに対しては，検索結果中のトピックに基づいた検索結果のクラスタリングにより検索結果を構造化し，絞り込み検索を支援する手法を提案する．提案手法の主な特徴は，トピック分類性に富んだ固有名詞に注目した点，および検索結果中での重要性と検索条件との関連性に基づく基準により検索結果のクラスタリングに効果的な固有名詞を選択する点であり，これにより，検索結果をクラスタ化するだけでなく，それぞれの

クラスタが何のトピックについて示しているかをわかりやすく提示する．

また，二つ目のケースに対しては，検索結果中に含まれる主要なトピックおよびトピック間の関係，トピックと個々の文書の関係を明らかにする検索結果の構造化により，ユーザの情報取得を支援する手法を提案する．提案手法の主な特徴は，個々の検索結果の文書をノード，文書間の関係をエッジで表現した「文書集合グラフ」により検索結果の集合を表現し，そのグラフ構造とグラフ構造中の各ノードの中心性を利用して，検索結果中のトピック間の関係などを表現するトピック構造を明確化しようとする点である．これにより，検索結果の文書をクラスタ化するだけでなく，「トピック間の関係が知りたい」や「特定のトピックを最も良く表現する文書を取得したい」という要求にも対応できる検索結果の分析を実現する．

目次

図目次

表目次

第 1 章

緒言

コンピュータネットワークの発展により，アクセス可能な情報は増大し，効率的な情報アクセス手段へのニーズが高まっている．テキストのみに注目しても情報量の増大はとどまることを知らず，WWW に存在する文書は 100 億を越え [16]，また企業や官公庁をはじめとする組織内に存在するテキストも同様に膨大となっており，これらの情報を取捨選択し，効率よく利用することがますます重要となっている．

テキスト情報に対する一般的なアクセス手段としては，Google[88] や Yahoo![93] 等の Web サーチエンジンに代表されるキーワードベースの検索システムが一般的に用いられている．このシステムを利用する場合，ユーザは適切なキーワードをシステムに入力し，そのキーワードを含む検索結果のリストを取得する．そして，そのリストの中から，個々の検索結果のタイトルやスニペットの情報，場合によっては本文を閲覧する事で，真に欲しい情報を選別し，取得する．

しかし，最後の検索結果のリストから真に欲しい情報を選別する作業は，しばしばユーザにとって困難かつ時間がかかる作業となる．そのような状況に陥る主要なケースとして，以下に示す二つのケースが考えられる．

一つ目のケースは，特定の文書を求めて検索を行うが，検索条件があいまいとなってしまう場合である．検索条件に指定されたキーワードが多義性を持っていたり，ユーザが明確な検索条件を作成できないという原因が考えられる．このような場合には，ユーザが求める情報は大量の検索結果の中に埋もれてしまい，その中からユーザは所望の情報を探さなければならない．

また，別のケースとしては，元々の目的が検索結果全体から，概要や特徴的な情報を抽出しようとする場合である．このような場合にも，ユーザは所望の情報を取得するために，膨大な検索結果を閲覧しなければならない．

このような状況を改善するため，分類やクラスタリングを利用し，検索結果の構造化を行う手法が研究されている．これらの研究では，検索結果の構造化により検索結果の全体像を明確化することで，ユーザが目的の文書へ到達し易くなる事を目的としている．

このうち分類を利用する手法は，事前に定義したカテゴリ構造に対して，個々の検索結果を割り当てる手法である．分類処理の概要を図 1.1 に示す．図に示すように処理対象の文書群に加えて，あらかじめカテゴリが用意されている場合に，処理対象の文書をいずれかのカテゴリに割り当てる処理である．この処理を行う事により，上記で示す最初のケー

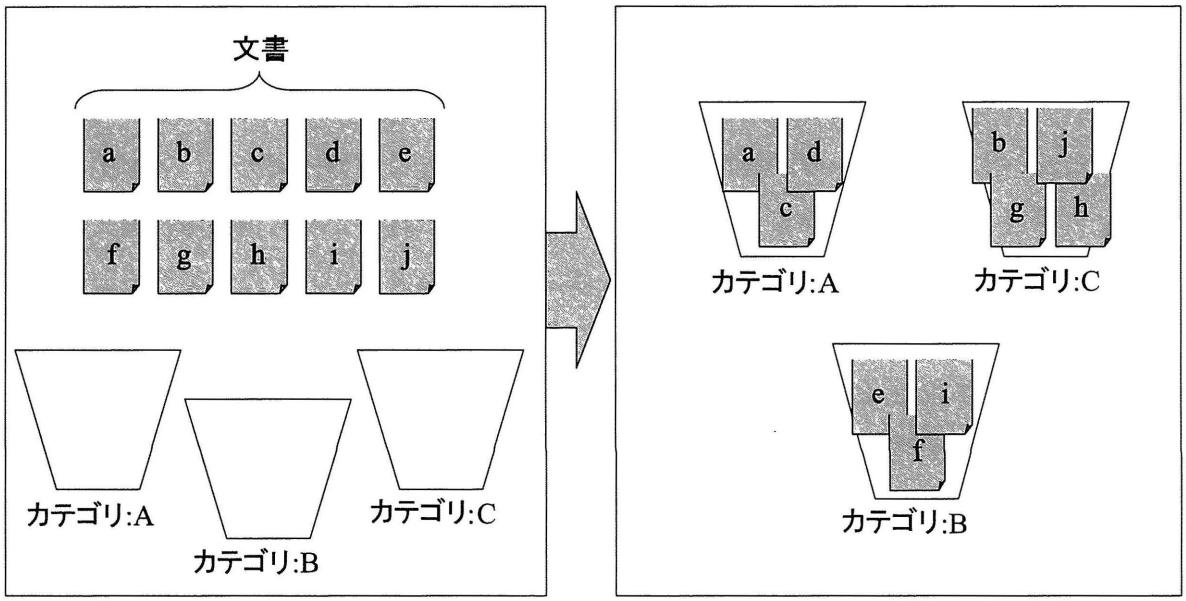

図 1.1 分類処理の概念図

スにおいて，ユーザは所望の結果が含まれると考えられるカテゴリを選択する事で，目的を満たす結果に容易に到達可能となる．しかし，カテゴリ構造の構築や，文書をカテゴリ構造への割り当てるために必要となる各カテゴリの特徴を表現するデータの作成は非常に時間のかかる作業であり，多様なトピックを対象とした検索に柔軟に対応するためには問題も多い．

一方，クラスタリングを利用する手法は，検索結果を類似した文書ごとにクラスタ化する手法である．処理の概要を図 1.2 に示す．図に示すように，文書群が与えられた場合に，それをいくつかのクラスタに分割する．この手法によると，上記の最初のケースにおいては，所望のトピックを含むクラスタを特定することで，容易に目的を満たす文書を見つけられるというメリットがあり，また後者の場合にも，個々のクラスタを閲覧する事で，検索結果中に存在するトピックの概要が把握できるというメリットがある．

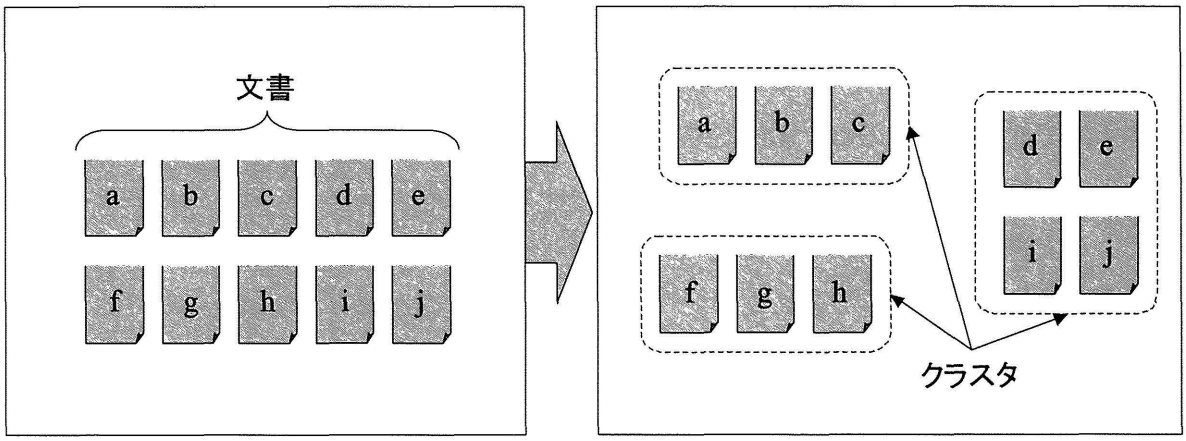

図 1.2　クラスタリング処理の概念図

しかし，一般にクラスタリングでは，個々のクラスタがどのような情報を含むかという事は考慮されておらず，これを行うクラスタへのラベル付けは困難なタスクであると言われている [8]．この問題により，上記に示したクラスタリングのメリットが十分生かせないのが現状である．さらに，クラスタリングは，文書のクラスタ化のみを行う手法であるため，文書集合中のトピックの関係や，特定のトピックに関する文書を探すという目的には不十分な点も多い．

以上で示したように，キーワードベースの検索システムにおいて，所望の情報を取得するために，大量の検索結果を選別，分析しなければならないという問題に対しては，十分な解決策が存在してないのが現状である．

本論文の目的は，上記に示した問題に対する解決策として，検索結果中に含まれるトピックに注目し検索結果を構造化する手法を提案する事にある．以下に個々の研究内容およびそれぞれの適用範囲について示す．

研究内容 1：固有名詞を利用した検索結果クラスタリング

本研究では，目的とする文書は少ないが，検索条件のあいまいさにより，検索結果の中に所望の情報が埋もれてしまうと言う問題に対して，検索結果中のトピックに基づいた検索結果のクラスタリングにより検索結果を構造化し，絞り込み検索を支援する手法を提案する．

提案手法の主な特徴は，トピック分類性に富んだ固有名詞に注目した点，および検索結果中での重要性と検索条件との関連性に基づく基準により検索結果のクラスタリングに効

果的な固有名詞を選択する点であり，これにより，検索結果をクラスタ化するだけでなく，それぞれのクラスタが何のトピックについて示しているかをわかりやすく提示できる．

本手法で対象とするのは，新聞記事，ブログ記事のように，固有名詞がトピックを見分ける上で有益な意味を持ち，かつ一つのテキストが特定のトピックに関する内容で構成されている文書である．ただし，複数のトピックを含む文書であっても，個々のトピックに関係する部分文書が特定できる文書については，部分文書を個々の文書とみなすことで本提案手法を適用する事が可能である．

研究内容2：グラフ分析を利用したトピック構造マイニング

本研究では，検索結果全体から特徴的な情報や概要を抽出したいと言う場合に，検索結果中に含まれる主要なトピックおよびトピック間の関係，トピックと個々の文書の関係を明らかにする検索結果の構造化により，ユーザの情報取得を支援する手法を提案する．

主な特徴としては，検索結果中の個々の文書をノード，文書間の関係をエッジで表現した「文書集合グラフ」により検索結果の集合を表現し，そのグラフ構造とグラフ構造中の各ノードの中心性を利用して，検索結果中のトピックやトピックの関係を表現するトピック構造を明確化しようとする点である．これにより，検索結果の文書をクラスタ化するだけでなく，「トピック間の関係が知りたい」や「特定のトピックを最も良く表現する文書を取得したい」という要求にも対応可能な検索結果の分析を実現する．

さらに，近年ニュース記事やブログ記事のようにタイムスタンプ付きの文書が多く普及しており，このような文書を処理するには，タイムスタンプを考慮した文書の分析が必要であると考え，上記に示した「文書集合グラフ」構築時に，文書間のタイムスタンプの近さを考慮する手法を提案している点も特徴の一つである．

本手法で対象とするのは，新聞記事，ブログ記事のように，一つのテキストが特定のトピックに関する内容で構成されている場合である．ただし，複数のトピックを含む文書であっても，個々のトピックに関係する部分文書が容易に特定できる文書については，部分文書を個々の文書とみなすことで適用する事が可能である．また，検索エンジンが検索結果のスニペットとしてKWIC(Keyword in context: 問い合わせキーワードの前後を含む部分文書) を返却する場合には，個々の検索結果を特定のトピックを持った文書とみなす事ができるため，提案手法を適用する事が可能となる．また，タイムスタンプを利用する手法は，上記を満たす文書が，タイムスタンプを持つ場合に適用できる．

本論文の構成

本論文の構成を以下に示す.

2 章では，まず本論文で提案する提案手法に関係する分類およびクラスタリングを用いた関連研究について示す．次に，提案手法 1 に関連する固有名詞抽出に関する関連研究について示す．その後，提案手法 2 に関連するグラフ分析を用いたテキスト分析手法および時間的近さを利用したテキスト分析手法に関する関連研究を述べ，最後にこれらの関連研究との関係を示しながら，本研究で提案する手法の位置付けを明確にする.

3 章では，研究内容 1 の固有名詞を利用した検索結果クラスタリングについて示し，IREX のテストコレクションである新聞記事コーパスを利用した評価および本提案手法を利用した応用事例を提示することで，提案手法の有効性を示す.

4 章では，研究内容 2 のグラフ分析を利用したトピック構造マイニングについて示し，新聞記事コーパスおよび NTCIR の Web 文書コーパスを利用した評価および検証結果を示し，提案手法の有効性を示す.

5 章にて本研究の成果をまとめ今後の課題について述べる.

第 2 章

関連研究

本章では，まず検索結果等の文書集合を構造化する手法として，分類およびクラスタリングを利用した研究について示す．次に，研究内容 1 に関連する固有名詞抽出に関する関連研究を示し，その後に，研究内容 2 に関連するグラフ構造を用いテキストの分析を行う関連研究および時間情報を利用しテキスト中のトピックの分析を行う関連研究を示す．最後に，これら関連研究と提案手法の位置付けについて示す．

2.1 分類に関する研究

分類に関する手法は，データの特徴を元にデータの集合をあらかじめ規定されたカテゴリに分類するアルゴリズム [62][41][5][75][51] を利用し，検索結果等の文書集合をカテゴリ体系に分類することで，ユーザの情報検索を支援する．

Open Directory Project[92] や Yahoo! Directory[94] 等において，大規模なカテゴリ構造と，個々のカテゴリを特徴付けるデータが存在し，ユーザのニーズも高いことから，Web ページの分類に関する研究は盛んに行われている．Dumais[8] らは，SVM を利用し，Web 文書を Web ディレクトリに分類する手法を提案している．この研究では，分類を行う際に，ディレクトリの階層構造を利用する事で，利用しない場合と比較して，精度が向上し，計算量が低減できる事を示している．Liu ら [40] は，Web 文書を Web ディレクトリに分類する際に，トレーニングデータの少ないカテゴリの分類精度が低下することを避けるため，木構造に基づき，カテゴリの特徴を伝播させる事で，分類精度を向上させる手法について示している．Sun[59] らは，Web 文書の分類において，サイトのタイトルやリンク元のアンカーテキストを重視する事で分類精度が向上する事を示している．Kwon ら [36] は，Web サイトを分類するために，Web サイトの文書群から，そのサイト

を良く表現するWeb文書を選択し，それをK-NN[5]を利用してWebディレクトリに分類し，その結果に基づきWebサイトの分類を行う手法を提案している．

また，Web文書に限らない手法としては，以下の手法が挙げられる．仲川ら[84]は，より広いトピックの文書の分類に対応するため，複数のカテゴリ情報(一階層の分類情報)を切替えて利用する手法を提案している．Fukumotoら[13]は，大規模なテキストを分類する場合に，比較的軽く動作するNaive Bayesと，高次元のデータを扱えるSVMを組み合わせたテキスト分類手法を提案している．提案では，元々のトレーニングデータを利用したクロスバリデーションでNaive Bayesでは分類できなかった文書を，SVMのトレーニングデータとし，Naive Bayesで正しく分類できないデータをSVMで分類するとしている．Siersdorferら[57]は，トレーニングデータが大量に存在する場合に，トレーニングデータを複数に分割し，個々のトレーニングデータセットで学習した分類器の結果を統合する事で分類精度が向上する事を示している．

2.2 クラスタリングに関する研究

検索結果等の文書集合を構造化して提示する手法のうち，カテゴリ体系等の事前知識を必要としない手法としてクラスタリングを利用する手法がある．

いわゆるクラスタリングを利用する手法は，個々の文書をキーワードベクトル[76]で表現し，K-Means[18]や凝集法等のクラスタリングアルゴリズムを利用し，文書集合を構造化しようと言う手法である．しかしながら，検索の支援に利用するためには，個々のクラスタがどのような文書を含んでいるかを示すラベルが重要である事が認識され，近年，クラスタリング処理を文書集合から重要キーワードを抽出するタスクとみなす手法が提案されている．この手法では，抽出された個々のキーワードを含む文書群が個々のクラスタとみなされる．

以下では，キーワードベクトルを利用する前者の手法を“文書指向のアプローチ”，後者の利用しない手法を“ラベル指向のアプローチ”とし，それぞれのアプローチについて関連研究を示す．

2.2.1 文書指向アプローチ

文書指向のアプローチでは，それぞれの文書をキーワードベクトル[76]で表し，ベクトル間の類似度を元に，K-Means[18]やWard法[65]をはじめとする凝集法，Single-Pass法[22]，Fractionation法[7]等のクラスタリングアルゴリズムを利用することで，類似し

た文書をクラスタ化する.

Cutting らは, 検索結果等の大量文書を効率的に参照する手段として, Scatter/gather[7][21] を提案している. 本手法では, キーワードベクトルの類似度を元に, 高速にクラスタリングできる Fractionation 法 [7] を用いて, 文書集合をクラスタリングして提示, ユーザが選択したクラスタの文書を対象に, 再クラスタリングを行う. この繰り返しによって, 所望の文書に到達することを支援するという手法である. また, Leuski[39] は, キーワードベクトル間の類似度を元に, 凝集法によって検索結果をクラスタリングする手法について検討している. プロトタイプシステムを用いた評価では, 検索結果のリストを提示するだけのシステムと比較し, ユーザが所望の文書に到達するまでに閲覧する文書数が低減していると報告している. Haveliwa[20] らは, 検索時にクラスタリングを行うのではなく, あらかじめ検索対象の文書集合をクラスタリングしておき, それを検索結果の提示に利用する手法を提案している. また, 検索時の処理時間を考慮し, 検索結果のタイトルやスニペットのみを利用してクラスタリングを行う手法も提案されている [37][49].

また, Web 文書を対象とした場合に, 文書間の類似度に加えてハイパーリンクの情報も利用する手法も提案されている. He[24] らは, Web ページの集合から, 含まれるトピックを分析するため, キーワードベクトル間の類似度に加えてサイト間のリンク構造を利用することで, Web ページを分類し, 分類した結果からトピックを抽出する手法について提案している. また, Wang らもリンクのみを利用した Web 文書のクラスタリング [63] およびリンクと文書内容の両方を利用した Web 文書のクラスタリング [64] について示している.

2.2.2 ラベル指向アプローチ

ラベル指向のアプローチは, 文書集合内のタームの出現状況等から特徴的なタームをラベルとして抽出, 文書集合とともに提示する手法である.

Zeng ら [70] は, TF-IDF[53] 等のラベルに関する特徴と, そのラベルを含む文書集合の特徴の両方を用いて, 有益なラベルを選択する手法を提案している. また, Sakai ら [52] は, ラベル選択の基準として, TF-IDF に「絞り込み語に有効な語は検索結果中に分散している」との仮定に基づく値を考慮した指標を提案している. また, Ohta ら [47] は, TF-IDF に語の出現する文書のランキングおよび個々のテキスト中でのタームの出現位置を加味した指標を提案している. また, その他にも TF-IDF を基準として採り入れている研究は多い.

また，Zamir らの手法 [69] は，検索結果中のフレーズに注目した手法である．この手法では，suffix-tree[17][66] を利用して検索結果のスニペット中のフレーズを高速に集計し，各フレーズを含む文書の数と，各フレーズの構成語数を元にクラスタのラベルとなりうるフレーズを抽出し，クラスタのラベルにするという手法である．Ferragina らの手法 [11] や Zhao[71] らの手法も同様に検索結果のスニペット中のフレーズに注目した手法である．

また，Kummamuru ら [35] は，抽出されるラベルリストを文書集合のタクソノミと見なし，文書の網羅性やラベルリストのコンパクトさ等のタクソノミらしさを重視したラベルリスト抽出法を提案している．

さらに，技術の詳細は明らかにされていないが，上記で示した TF-IDF を用いた技術に近いものとして，Vivisimo[90] や，Mooter[89] が，Web 上の検索エンジンとして実用化されている．

2.3 固有名詞抽出およびそれを利用したテキスト分析に関する研究

固有名詞を抽出する技術として，固有表現抽出 [25][26][3][14][48][30] があげられる．固有表現抽出は，1990 年代に，MUC[15] において生まれたタスクであり，新聞記事等のテキストから，内容を理解するために重要であると考えられる人名や組織名等の固有名詞と金額表現や割合表現等の数値表現を抽出するタスクである．現状で高精度な固有表現抽出が出来るのは，新聞記事などの限定された分野のテキストに限られるが，より多くの分野の語彙を抽出しようという取り組みは盛んに行われており，抽出するカテゴリ数を 200 に増やすための研究 [54] や，メール等の形式的でないテキストからの固有表現抽出に関する研究 [45]，バイオインフォマティクスの分野で蛋白質や遺伝子の名前を抽出しようとする研究 [28] も行われている．

また，これに関係する研究として自然言語のテキストや HTML 文書を解析することで，特定のカテゴリに属する語彙を抽出しようとする取組がある．Pasca ら [50] は，自然言語中でカテゴリを表現する語に続いて，そのカテゴリのインスタンスである語が出現するパターン (例：「カテゴリ語」 such as 「語 1」, ... 「語 n-1」, and 「語 n」) を用いて，大量のテキストを解析することでさまざまなカテゴリの語彙を獲得できる事を示している．また Shinzato ら [55][56][78] は，HTML 文書中に存在する繰り返し構造に注目し，同じカテゴリに属する同位語を抽出する手法を提案している．さらに，山口 [85] らは，検索キーワードに注目し，同じ語と組み合わされて利用される検索キーワードは同位語である

と言う仮説の元，検索キーワードのログから同位語を抽出する手法を提案している．これらの手法により，各カテゴリ毎の辞書を構築する事ができれば，上記で示した固有表現抽出と組み合わせることでより多様なテキストから，多様な固有名詞を抽出することが可能となる．

一方，このようにして抽出される固有名詞を利用する手法としては，質問応答が挙げられる．この分野の研究では，人名を求める質問には人名を回答すると言うように，質問のタイプに応じた解答の候補を抽出する手法として固有表現抽出が利用されている [77][38][58].

また，以下に示す研究では，固有名詞がトピックを見分けるために有益であることを示している．Yang ら [68] や Kumaran ら [32] は，ニュースストリームからトピックを検出する場合に，固有名詞を利用する事で，重複したトピックの検出が低減した事を示しており，戸田ら [81] は，日本語のニュース記事の検索結果をクラスタリングする場合のキーワードベクトルの各次元に形態素解析の結果得られる名詞等をそのまま利用する場合と，固有表現抽出により抽出された固有名詞を利用した場合について比較実験を行い固有名詞を利用する事でクラスタリング精度が向上する事を示している．さらに，木村ら [29] は，同姓同名が存在する人名で Web 検索を行った結果を，個々の人物ごとにクラスタリングするタスクにおいて，固有名詞を利用することで精度が向上することを示している．

さらに，Zhang ら [72] はテキスト中に存在する固有表現のうち，最もテキスト中のトピックを表現する固有表現 (群) を抽出するため，固有表現の出現する場所や，回りのコンテキストをもとに機械学習を利用する手法について示している．

2.4 グラフ分析を利用したテキスト分析に関する研究

近年，文書やセンテンス等の言語的な要素間の関係を元にグラフ構造を構築し，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 (文書やセンテンス等) のランキングやいくつかの要素を抽出する手法が提案されている．

Mihalcea[42][43][44] らは，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 アルゴリズム [4] が有益に働くことを実験的に示した上で，重要文抽出とキーワード抽出の実験を行い，有益な結果を得たことを報告している．Erkan[10] もまた，文書集合から，重要なセンテンスを抽出するタスクにおいて，各センテンス間の類似度を元にセンテンスをノード，センテンス間の関係をエッジとしたグラフ構造を考え，そのグラフ構造中の中心

性を用いた手法を提案しており，既存手法との比較でより高い精度を示すことを報告している．さらに，Kurland ら [33] は，文書間にリンクのない文書集合において PageRank を計算するため，言語モデルを用いたグラフ構成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．さらに，Kurland は，言語モデルを用いたグラフ構造を利用して，HITS アルゴリズム [31] により検索結果の再ランキングを向上させる手法 [34] も提案している．

2.5 時間的近さを考慮したテキスト分析に関する研究

文書の発行されたタイムスタンプを意識し文書集合のトピックを分析する取り組みとして石川らの研究 [74] があげられる．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トピックを特定するために，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，重要と思われる最新のトピックのみを選択的にクラスタリングする手法を提案している．

Yang ら [67] 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．

また，Cui ら [6] 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．

2.6 提案手法の位置付け

本研究の研究内容 1 は，検索結果のクラスタリングを検索結果から重要キーワードを抽出する事で実現しようと言う点で，2.2.2 に示したラベル指向アプローチのクラスタリングの一手法であると言える．しかし，ラベル候補の抽出に，固有表現抽出を利用し，トピック分類性に長けた固有名詞をラベル候補として利用する点，そのラベル候補の中から，検索結果中での重要度と検索条件との関連性に基づいた基準を利用して，検索結果の分類に効果的なラベルを選択しする点で従来手法と異なり，よりわかりやすい検索結果の概観および高精度の絞り込み検索を実現する．

また，本研究の研究内容 2 は，キーワードベクトルを利用した文書間の類似性を元に，文書集合の分析を行う点で，2.2.1 で示した文書指向アプローチのクラスタリングと関連

するが，文書集合をグラフ構造で表現し，相互に関係が強い文書の集まりをクラスタとして選択的に抽出する点，また単にクラスタを抽出するだけでなく，複数のクラスタ間の関係や，それぞれのクラスタの中心的な内容に対する各文書の位置付けを明確化する点で異なる．これにより，文書集合の全体像把握および内容理解を支援する．また，本研究は，グラフ構造を考え，各ノードの中心性を考慮する点で，2.4 で挙げたグラフ構造を利用する手法に関係するが，本研究で提案する手法は，中心性スコアをグラフ構造とともに利用する事で，文書集合中に存在するトピックを特定し，複数のトピック間の関係等を抽出する点で異なる．さらに本研究では，タイムスタンプ付きの文書を扱うために，文書間の類似度として，文書内容の類似度に加え，時間的近さを考慮する手法についても分析している．この点で 2.5 で示した時間的近さを考慮したテキスト分析を行う手法に関連するが，類似度算出をグラフ構造構築時に利用するという点で従来手法とは異なる．

第 3 章

固有名詞を利用した検索結果クラスタリング

3.1 はじめに

キーワードベースの検索システムを利用する際に問題となる点として，検索条件があいまいな場合に，検索結果が膨大となり，所望の情報が検索結果中に埋もれてしまうという点がある．原因としては検索条件に指定されたキーワードが多義性を持っていたり，ユーザが明確な検索条件を作成できないという場合が考えられる [2].

この状況を改善する策としては，検索要求をより明確にするため，ユーザにより多くの検索キーワードを指定させる事が考えられる．これにより，検索結果は絞り込まれ，検索結果のランキングもよりユーザの要求にあう形になると考えられる．しかし，検索システムのユーザは 3 語以上のキーワードを入力する事はほとんどない [1] という調査結果もあり，一般的なユーザにより多くの検索キーワードを指定させることは困難である．

このような問題を解決するため，検索結果のクラスタリングを行う手法が提案されている．これらの手法の目的は大きく別けて 2 つである．一つは，明確な検索目的を持っているユーザを支援するためであり，ユーザは最も適切なクラスタを選択することで，所望の情報に効率的に到達できる．もう一つは，検索目的が明確でなく，比較的あいまいな検索条件で検索を行う場合の支援である．この場合，ユーザはクラスタの一覧を参照して，検索結果の概要をつかむ事が可能となり，検索要求の明確化につながる事も考えられる．

また，二次的な効果であるが，元々検索している事と関連するが，予期しない情報を発見するという効果もある．例えば，アメリカ合衆国の大統領選挙について検索をするために，「大統領選挙」というキーワードで検索を行うと，アメリカ合衆国の大統領選挙に関

するクラスタの他に，別の国の大統領選に関するクラスタが存在し，他の国の大統領選について知ることができる．

しかしながら，上記で述べたようなメリットを実現するには，単に検索結果をクラスタ化するだけでなく，個々のクラスタを説明するラベルが十分わかりやすいことが重要となる [69]．このような点を考慮し，いわゆるクラスタリングアルゴリズム ([18][65][22][7] 等) を用いる手法 ([21][39][24] 等) ではなく，検索結果のクラスタリングを検索結果からの重要ターム (単語やフレーズ) の選択とみなした手法 ([69][70][47][11][35] 等) が提案されている．

本研究では，このアプローチをさらに進め，クラスタリングを検索結果のインデクスを作成するタスクと見なす．ここで言うインデクスとは，構造化したラベルリストの事を示しており，これによりユーザはより簡単にラベルを閲覧し，検索結果の概要を把握できる．これを実現するため，本研究では 2 つの点を提案する．

一つは，ラベルの候補となるタームの抽出に固有表現抽出 [15] を利用し，固有名詞の抽出を行う事である．これは文書のトピックを特徴付ける情報として固有名詞が有益であると考えたためである．また，固有表現抽出によって抽出されたタームに付与される種類 (人名，地名等) をラベルの提示時に利用する事で，同種のラベルをまとめ, 構造化した形で提示できる．二つ目は上記で抽出したラベル候補からユーザに提示するラベルを選択する新たなラベル選択基準である．この基準では，検索結果内でのタームの重要性と，検索条件との関係性に基づき各タームを評価し，これにより絞り込み検索に有益であると考えられるラベルの抽出を可能とする．

以下，本章の構成を示す．3.2 では，既存手法の問題点を明確化し，それに対する本研究のアプローチについて述べる．3.3 では，前記のアプローチに基づく提案手法について示す．3.4 で新聞記事コーパスを利用した評価について示し，3.5 では，提案手法をニュース記事，ブログの記事検索システムおよび話題提示システムに適用した応用例を示す事で，提案手法の有効性を示す．最後に 3.6 でまとめる．

3.2 従来技術の問題点とアプローチ

3.2.1 問題点の明確化

本節では，ラベル指向のクラスタリングを利用した Web サーチエンジンとして公開されている Vivisimo[90]，Mooter[89] を例に挙げ，既存技術の問題点を示す．

それぞれのシステムは日本語には対応していないため，英語の検索条件を用いて利用

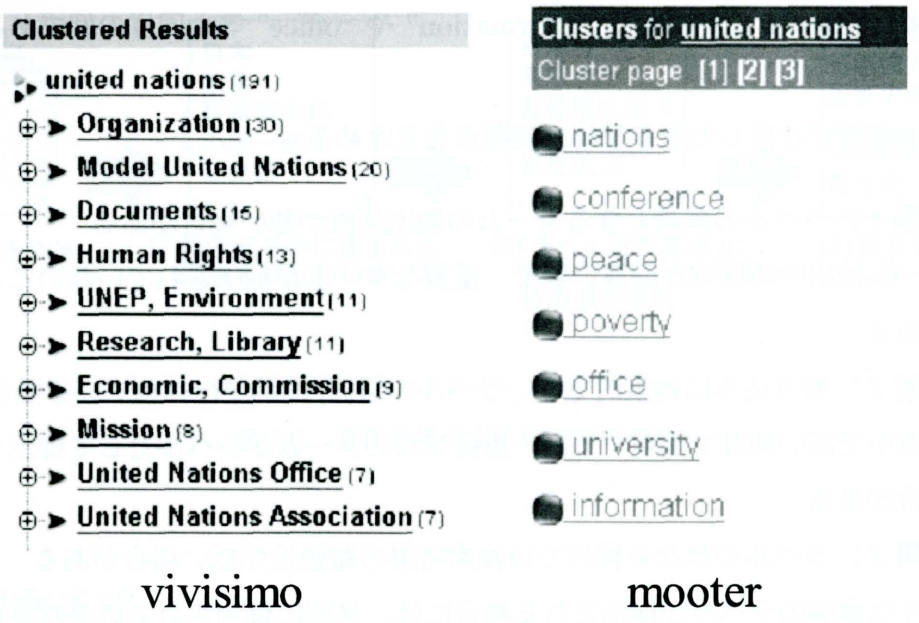

図 3.1 既存システムのラベルの例

した.

Vivisimo および Mooter で “united nations” というキーワードで検索した場合に出力されるラベルを図 3.1 に示す. Vivisimo の例を見ると, “Organization” や “Documents” 等の一般的なタームと, “Model United Nations” や “UNEP” 等の具体的な対象を示すタームがラベル中で混在していることがわかる.

このうち, 一般的なタームである “Organization” および “Documents” について, それぞれ関連する検索結果中の文書を参照し, ラベルとして存在している理由について考えた.

まず, “Organization” について見ると, 多くの場合, 様々な機関の名称中に存在し, それらの機関の名称の末尾から “Organization” というタームだけが抽出されていることがわかる. これにより, 検索結果中での出現頻度が高くなりラベルとして提示されていることが考えられる. Mooter で出力される “Nations” や “Conference” 等も同様な例であると考えられる.

また, “Documents” というラベルに関連する文書には何らかの文書情報もしくは文書情報へのリンクが存在することが想定されたが, このラベルに関連付いたサイトは国連に関係する機関のトップページが多く, 想定した情報はあまり存在しない. このようなタームは単に文書に出現する頻度が高いためにラベルとして提示されているものであると

考えられる．Mooter の例にある “information” や “office” 等も同様の例であると考えられる．

以上に本研究で注目した既存技術の問題点をまとめる．

- **問題** *1*： ラベルの候補となるタームの抽出品質が悪い場合がある
  ターム抽出時の区切り間違い等で，重要なタームが抽出されていないことがしばしばある．
- **問題** *2*： 絞り込みに効果的でないラベルの選択/提示が行われている場合がある
  文書中で高い頻度で出現するが，重要でないタームがラベルとして提示されている場合がある．
- **問題** *3*： ラベルの単なる羅列では検索結果の概観性が低い場合がある
  様々な意味のラベルが提示される場合には，単純に羅列されているだけでは，検索結果の全体像が掴みにくく，絞り込み語を探す場合にも探しにくい．

3.2.2 アプローチ

本節では前節で指摘した既存技術の問題点を解決するための本研究のアプローチを示す．

上記の**問題** *1* に関して，本研究では，固有表現抽出 [15] を利用して抽出できる固有名詞をタームとして利用することを提案する．固有表現抽出では，文書中に含まれる「人名」，「組織名」，「地名」等の固有名詞を高精度で抽出できる[*1]．これらの固有表現は元々新聞記事などの文書中で「頻繁に重要になり，情報としての単位がはっきりしている」表現と定義されており [15]，検索においても重要な表現であると考えられる．

また，固有表現抽出によって抽出されたタームに付与される種類 (人名，組織名，地名等) をラベル提示時に利用する事で，同種のラベルをまとめた構造化されたインデクスを生成することが可能となり，**問題** *3* に対する解決策となる．

問題 *2* に関しては，ラベルとして有効な特性を検討した上で新たなラベル選択基準を提案する．このラベル選択基準に基づきラベルとなるべきタームを選択し，検索結果の絞り込みに有効なインデクスの提示を可能とする．

[*1] 固有表現抽出技術では，上記で示した固有名詞の他，「金額表現」や「割合表現」等の数値表現を抽出することが可能であるが，今回は固有名詞のみを利用する．

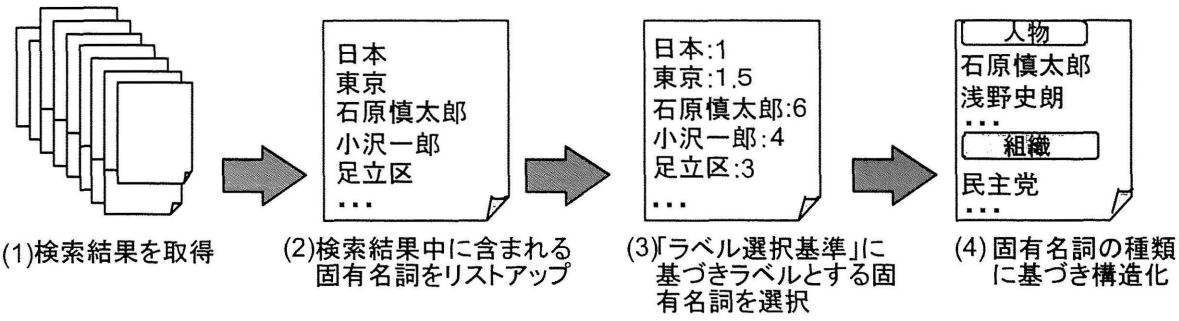

図 3.2 提案手法による検索時の処理

3.3 提案手法

3.3.1 手法概要

まず，文書を検索システムに登録する段階で，文書を全文検索用のインデクスに登録するとともに，固有表現抽出技術を用い文章中の固有名詞を抽出し，どの文書にどの固有名詞が含まれるかという情報をターム出現情報データベースに格納する．実際に検索要求が来た場合の処理については，概要を図 3.2 に示し，個々のステップについて以下に示す．

1. 検索結果を取得
   入力された検索キーワードをもとに全文検索システムに問い合わせ，検索結果の文書集合を特定する．
2. 検索結果中に含まれる固有名詞をリストアップ
   上記で取得した検索結果に含まれる固有名詞を，ターム出現情報データベースから取得しリストアップする．
3. 「ラベル選択基準」に基づき，リストからラベルとする固有名詞を選択
   リストアップされた個々の固有名詞を「ラベル選択基準」でスコアリングし，スコアが高いものから規定数分の固有名詞をラベルとして選択する．
4. 固有名詞の種類に基づきラベル群を構造化して，検索結果とともに提示
   上記で選択されたラベルを，固有表現抽出で付与された種類に基づきまとめ，提示する．

次に，上記で示した「ラベル選択基準」の詳細について示す．

3.3.2 ラベル選択基準

ここでは，検索結果中に含まれる固有名詞の中から，検索結果のクラスタリングに有益な固有名詞をラベルとして選択する「ラベル基準基準」について示す．

検索結果の中からラベルを選択する問題は，文書集合の中から重要語を選択する問題ととらえる事ができ，多くの関連研究では，TF-IDF[53] を用いた基準が利用されている．TF-IDF は，もともと文書検索における索引語の重み付けのために提案された基準であり，注目している文書中での出現頻度が多い事が重要という局所的重みと，文書集合全体で希少であるタームが出現する事は重要という大域的重みから構成される．ただし，TF-IDF による重要語選択では，過度に TF 項の影響が強くなる傾向が指摘されている [23]．また，文書集合からの語彙抽出に TF-IDF を用いる場合，TF を文書集合中でその語が出現した文書数で置き換える場合があるが，これも TF と同じ傾向を示す．

そこで本研究では，検索結果の概観や絞り込み検索の効率化に有効なラベルの要件を以下のように考え，これらに基づく新たな基準を提案する．

- 検索結果の中で当り前でなく，かつ希少過ぎないタームが重要
- 検索条件との関連性が高いタームが重要

最初の項目を具体化するには，検索結果中での頻度が過度に多くもなく少なくもないタームを評価する指標が必要となる．そこで，以下の式に示すように検索結果数 (ラベルを選択するために，全ての検索結果を解析せずに，検索ランキングの上位のものだけを解析する場合には，処理結果数) に対して 30〜40% 程度の出現頻度で値が最大となるような基準 (ラベル選択基準 1) を提案した．

$$\text{ラベル選択基準 } 1_i = DF_{R,i} \times \log(\frac{|R|}{DF_{R,i}}) \tag{3.1}$$

ここで，R は検索結果の文書集合 (もしくはラベル抽出の際に処理する検索結果の集合) であり，$DF_{R,i}$ は検索結果 R 中でターム i の出現する文書の数であり，“ラベル選択基準 1_i” は「検索結果の中で当り前でなく，かつ希少過ぎないタームが重要」という考えに基づくターム i の重みを示す．

上記に示したラベル選択基準 1 の関数形状を図 3.3 に示す．このグラフでは，横軸が検索結果中で該タームを含む文書がどの程度の割合であるかを示し，また，縦軸はラベル選択基準 1 による評価値である．つまり，左端は検索結果中に該タームがまったく含まれな

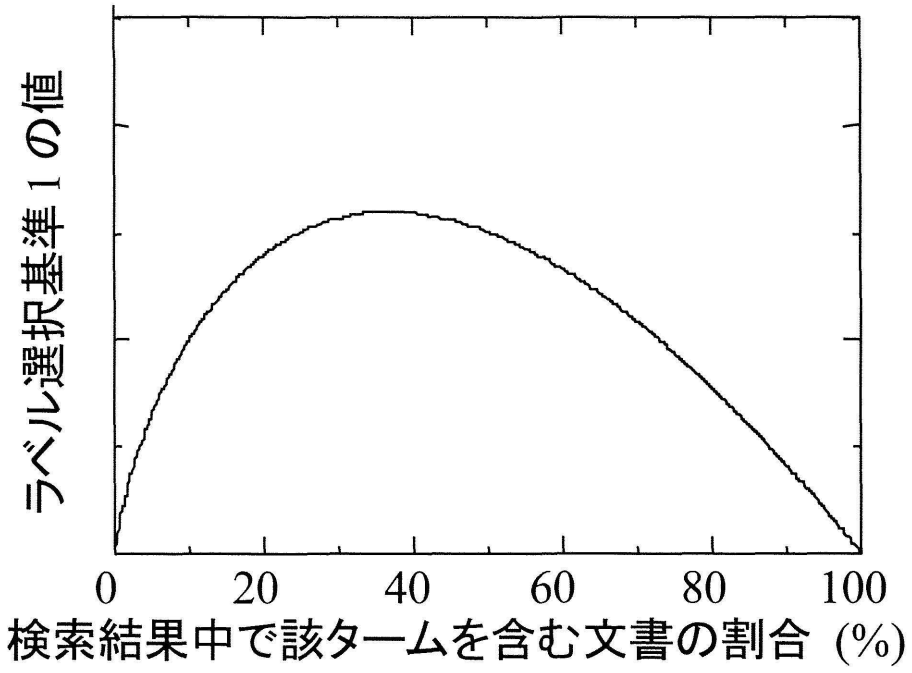

図 3.3 ラベル選択基準 1 の関数形状

い場合を示し，その場合の評価値は 0 である．また，右端はラベルが全ての検索結果中に含まれる場合を示し，その場合にも評価値は同じく 0 である．

一方，後者の項目「検索条件との関連性が高いタームが重要」を具体化するために，コーパス全体でのタームの出現率と，検索結果中でのタームの出現率の比を取ることで，検索条件と関連性の高いタームを特定できるのではないかと考えた．これは，コーパス全体での出現率より検索結果での出現率が高いタームは，検索条件と関連性が高いという考えに基づく．

これに基づく基準 (ラベル選択基準 2) は以下の式で算出できる．

$$\text{ラベル選択基準 } 2_i = \frac{DF_{R,i}/|R|}{DF_{D,i}/|D|} \tag{3.2}$$

ここで，D は文書コーパス全体の文書集合を表し，$DF_{D,i}$ は文書集合 D 中でのターム i の出現文書数を示す．また，“ラベル選択基準 2_i” は「検索条件との関連性が高いタームが重要」という考えに基づくターム i の重みを示す．

ただし，上記に示す式は，文書集合中でごく稀なタームが検索結果中に存在した場合，過度に高い評価を与える傾向があるため，その傾向を抑えた基準として以下も併せて提案する．

$$\text{ラベル選択基準 } 2a_i = \log \frac{DF_{R,i}/|R|}{DF_{D,i}/|D|} \tag{3.3}$$

$$\text{ラベル選択基準 } 2b_i = \sqrt{\frac{DF_{R,i}/|R|}{DF_{D,i}/|D|}} \tag{3.4}$$

実際にラベルの評価を行う際には，以下のように，上記で定義した二つの基準を組み合わせた式で各タームの評価を行う．

$$\text{ラベル選択基準}_i = \text{ラベル選択基準 } 1_i \times \text{ラベル選択基準 } 2_i \tag{3.5}$$

二番目の仮説を表現するために，式 (3.3) や式 (3.4) を利用する場合には，上記，ラベル選択基準 2_i を該当する式で置き換える．

3.4 評価

本節では，ラベル選択基準の評価について述べる．まず，3.3 で示した提案手法を実装したプロトタイプシステムおよび評価で利用するデータについて示す．その後，評価手法および評価結果について述べる．

3.4.1 評価用プロトタイプシステム

評価実施のため，以上で説明した手法に基づき，全文検索システム LISTA[19] および磯崎の手法 [26] による固有表現抽出ツールを利用し評価用のプロトタイプシステムを構築した．

LISTA のランキングアルゴリズムには，検索システム freeWAIS-sf で採用されている手法が利用されている [82]．システムは Web サーバ上に構築し，ブラウザを介してアクセスする．

図 3.4 にシステムの概要を示す．前処理として，検索システムに登録する文書から固有表現を抽出した後，全文検索用インデクスの生成およびターム (固有表現) の出現情報の算出を行う．検索時には，Web サーバ中に構築されたアプリケーションが全文検索機能および提案手法により構成されるインデクス生成機能にアクセスし，検索結果とインデクスを取得し，ユーザに提示する．

また，図 3.5 にユーザインタフェースを示す．図では，ユーザが入力した検索条件を元に，検索結果を表示した状態を示しており，右側に検索結果リストを提示，左側に複数の

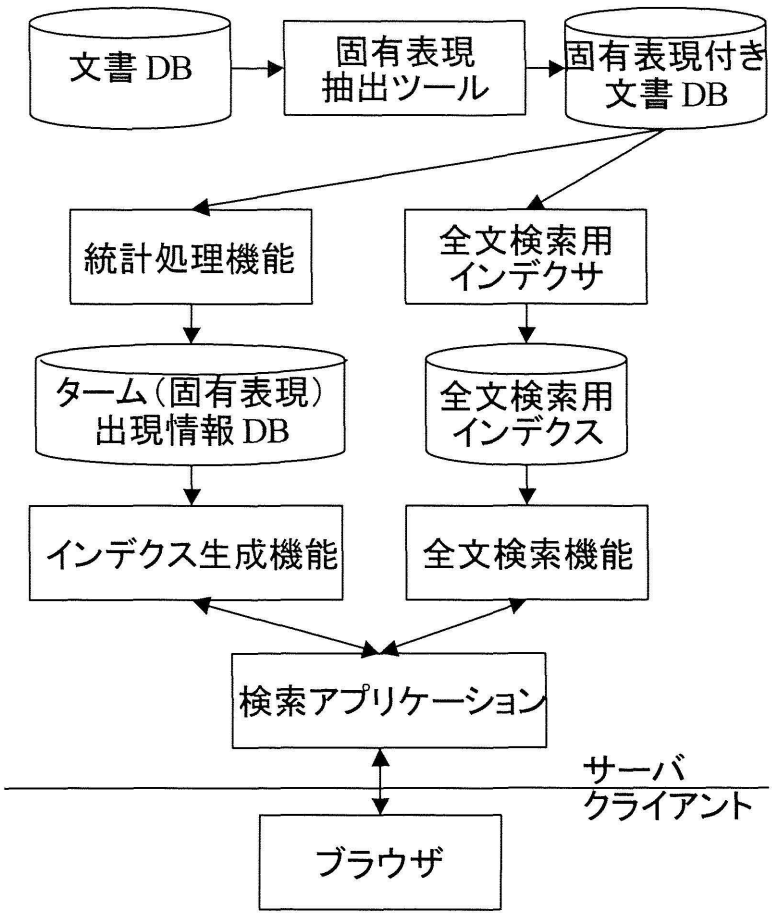

図 3.4 プロトタイプシステムの構成

ラベルから構成されるインデクスを提示している．ユーザは，従来システムのように検索結果リストから所望の文書を選択することに加えて，インデクスを参照することにより，検索結果を概観したり，インデクス中に目的のラベルが存在する場合には，それを選択することで容易に検索結果を絞り込むことができる．絞り込み検索は，既に入力済みの検索条件と選択されたラベルのタームを “and” で結合し，再検索することで行われる．

3.4.2 評価用データ

本評価では検索精度の評価を行うため，IREX(Information Retrieval and Extraction Exercise)[79] で利用されたテストコレクションおよび，検索課題，個々の検索課題に関する適合性判定結果を利用した．テストコレクションは，1994 年および 1995 年の毎日新聞記事であり，文書数は約 20 万件である．また，このテストコレクションに対して 30 の検索課題および各検索課題に対する適合性判定が規定されており，各検索課題について平均

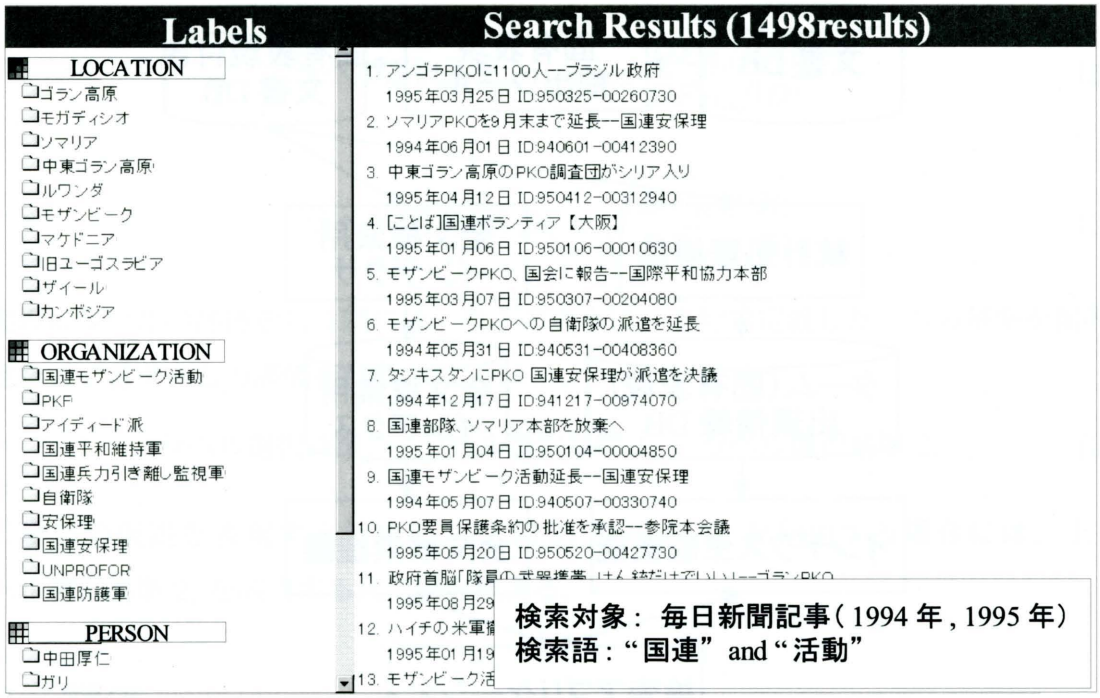

図 3.5 プロトタイプシステムのユーザインタフェース

約 100 件の適合文書が存在している (最小 29 件，最大 300 件).

評価用の検索条件の作成には，各検索課題について規定されている DESCRIPTION[*2]を利用した．これを形態素解析し，ストップワードを除去した後に “or” で連結し検索条件とした.

また，上記の適合文書中に含まれる個々の話題を対象に，絞り込み検索を行う場合に利用すると考えられるキーワードのリストとして「有効タームリスト」を作成した．作成手順としては，5 人の被験者に IREX で規定された検索課題および適合文書を提示，その適合文書の中から，「個々の話題の検索において検索条件として有効なキーワード」を選択してもらった．このうち 3 人以上の被験者が挙げたものを「有効タームリスト」に追加した．これをそれぞれの検索課題について作成した.

3.4.3 評価方法

ラベル選択基準を評価する場合に重要であると考えられるポイントは 2 点ある．1 点目は，ユーザが選択するラベルが提示できているかどうかであり，ラベルのわかり易さに関

[*2] 3 つまでの自立語から構成される検索要求の簡潔な表現である [79]

係する．もう一点は，その選ばれたラベルが適合文書と関係しているかどうかである．

この考えに基づき，本評価では，「ユーザが選択すると考えられるラベル」を選択して絞り込み検索を行った場合の検索精度を評価指標とした．システムによって抽出されたタームのうち，上記で示した「有効タームリスト」に存在するものを「ユーザが選択すると考えられるラベル」とした．以下に，評価の手順を示す．

1. システムに検索条件を入力し，検索結果とインデクスを取得する．
2. インデクス中の各ラベルを「有効タームリスト」と比較し，一致するラベルを「ユーザが選択すると考えられるラベル」として取得する．
3. 上記で取得した各々のラベルを用いて絞り込み検索を行い検索結果を取得する．
4. 上記で得たそれぞれの検索結果のうち，規定ランキング以上の文書について「判定対象」とし，その適合性判定を行う．

4 番目のステップの規定ランキングには，ユーザが文書を参照する件数ということで，10 とした．これは，ユーザが検索結果の 1 ページ目をすべて見ることを仮定した値である．またラベルの提示数は，XGA の解像度を持つディスプレイにおいて，1 画面で提示できるラベルの数ということで 20 とした．

今回の実験では，検索結果のうちランキング上位 n 件 ($n = 30, 50, 100, 200, 300, 500$) の検索結果をラベル選択に利用する文書として実験を行った．それぞれの条件で，処理する文書集合中に含まれるタームのうち，ラベル選択基準のスコアが高いものから規定数分 (今回は上述の通り 20 個) をラベルとした．

本評価においては，比較対象として，検索結果中の頻度を利用した方式，TF-IDF を利用した方式，TF-IDF を改良した方式である対数化 TF-IDF を利用した方式を利用した．また，提案手法，比較対象手法のいずれの手法を利用する場合にも，検索結果中での頻度が 1 であるタームはノイズである可能性が高いと考えラベル候補から除外している．

評価で利用する適合率は，各検索課題毎に得られた結果を元に以下の式で計算し，平均した値を利用した．この値は，提示したインデクスのラベルのうち，「ユーザが選択するであろうと考えられるラベル」を全て選択し，その場合に得られる絞り込み検索結果の，規定ランキング以上 (今回の規定ランキングは 10) の文書の適合率を示している．

$$\text{各検索課題における適合率} = \frac{\text{「判定対象」中の適合文書数}}{\text{「判定対象」の文書数}} \tag{3.6}$$

また，再現率については，各検索課題毎に得られた結果を元に以下の式で計算し，平均

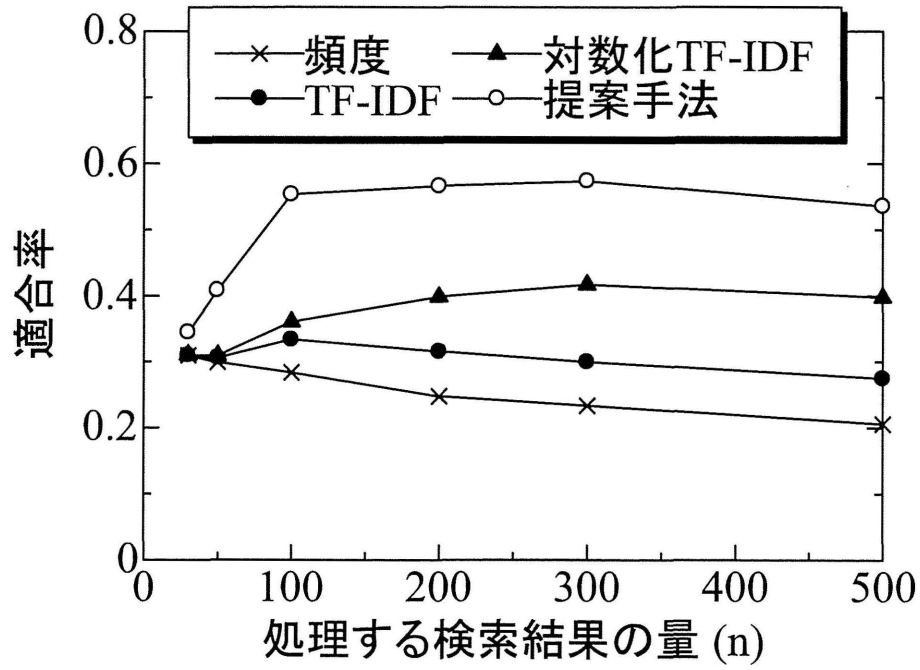

図 3.6 検索結果処理量と適合率の関係 (比較対象手法との比較)

した値を評価に利用した．この値は，インデクス中の「ユーザが選択するであろうと考えられるラベル」を全て選択し，得られた検索結果の規定ランキング以上 (今回の規定ランキングは 10) の文書を全て閲覧した場合に得られる再現率を示している．

$$\text{各検索課題における再現率} = \frac{\text{絞り込み検索で到達できた適合文書数}}{\text{適合文書数}} \tag{3.7}$$

3.4.4 評価結果

図 3.6 および 3.7 に，提案手法と比較手法の評価結果を示す．図 3.6 では適合率，図 3.7 では再現率について示す．横軸は，両方のグラフとも，ラベルを抽出するために処理した検索結果の件数を示す．縦軸はそれぞれ，適合率および再現率を示す．

まず，処理した検索結果の量と精度の関係について見ると，精度が上昇しているのは，適合率の場合，処理した検索結果が 100 件もしくは 200 件程度まで，再現率の場合も，50 件もしくは 100 件までである事がわかる．これ以上処理しても精度は向上していない．これは検索結果のうち過度にランキングが低い文書には適合文書がほとんど含まれないためであると考えられる．

次に，提案手法を比較対象手法と比較すると，処理した検索結果の量が少ない段階で

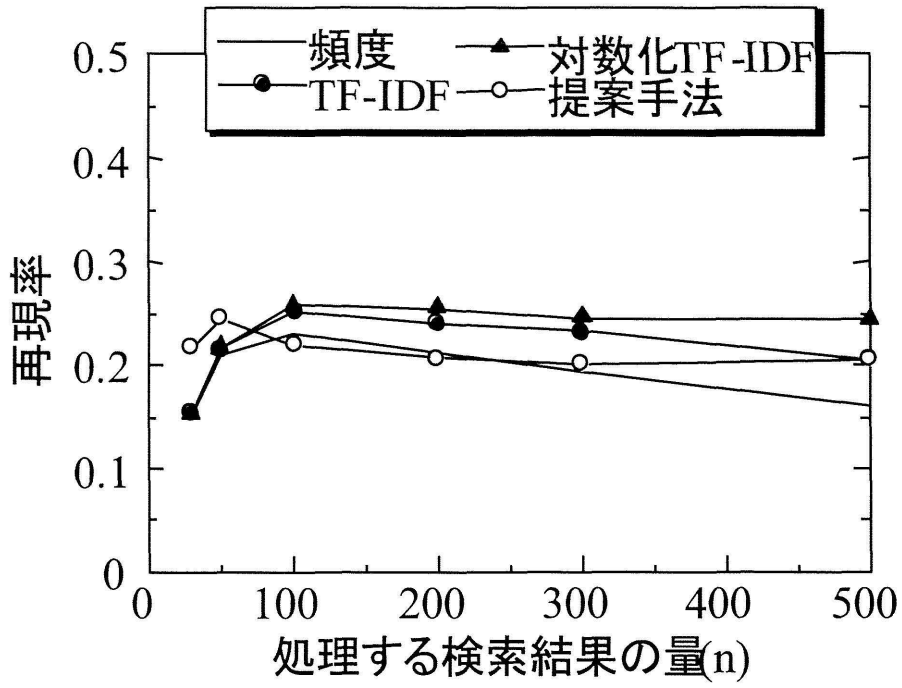

図 3.7 検索結果処理量と再現率の関係 (比較対象手法との比較)

は，適合率，再現率とも比較対象手法を上回り，$n = 50$ の場合には，比較対象手法が今回の評価で示した最高値に匹敵する再現率，適合率を示している．この特性は，より少ない計算量で高い精度を示すことができるという点で大きなメリットである．

また，全体的な傾向を見ると，提案手法の適合率は，処理する検索結果の量にかかわらず高い値を示しており，従来手法のうち最も高い適合率を示す対数化 TF-IDF とそれぞれの最高値を比較しても 15 ポイント以上高い値を示す．一方で，提案手法の再現率は処理する検索結果の量の増加に伴いやや低下し，同じ文書処理量で対数化 TF-IDF と比較した場合，最大で 5 ポイント程度の差がある．

この再現率低下の原因としては，提案手法の二つ目の仮説を説明する指標として式 (3.2) を利用した場合，ごく稀に出現する語彙が極端に高い評価を受ける事に原因があるのではないかと考えられる．つまり，多くの検索結果を処理した場合には，その検索結果に含まれる不適合文書中の希なタームが高く評価され，インデクス中のラベルとして提示されたのではないかと考えられる．実際，検索結果の処理数を増加させた場合，提案手法では「ユーザが選択すると考えられるラベル」の数が減少している ($n = 50$ の場合 1 検索課題あたり約 12 個が，$n = 200$ では約 8.5 個)．そこで，この極端な評価値の上昇を抑える式 (3.3) および式 (3.4) を利用した結果を図 3.8 および 3.9 に示す．

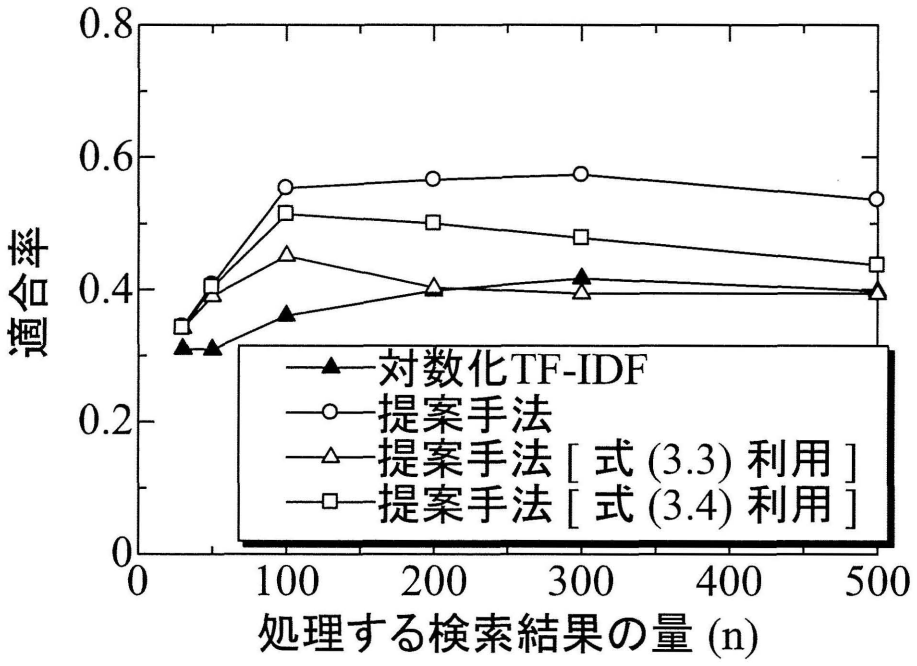

図 3.8 検索結果処理量と適合率の関係 (式 (3.3) および式 (3.4) を利用した場合)

提案手法の傾向として，検索結果の処理数が少ない場合にも適合率，再現率が高いという点は変わらない．また，検索結果の処理数を増加させた場合，式 (3.2) を利用した場合と比較して再現率は上昇し，比較対象手法との差は減少している．「ユーザが選択すると考えられるラベル」の数を見ると，$n = 200$ とした場合，一つの検索課題あたり式 (3.3) の場合 10.4 個，式 (3.4) の場合 9.8 個とやや増加しており，これが再現率上昇の要因であると考えられる．一方，適合率はやや低下している．これは個々のラベルが式 (3.2) を利用して抽出されたラベルと比較し，絞り込み検索時にノイズを混入させる事が原因である．これは，式 (3.3), (3.4) を利用することで，ラベルが多少一般的なものに変化したためであると考えられる．

以上より，提案手法の二つ目の仮説に関する式 (3.2), (3.3), (3.4) を使い分ける事は，必ずしも全ての状況に置ける精度が向上するものではないが，これらの式を使い分けることにより，適合率を重視したり，再現率を重視したりという調整が可能であることがわかった．

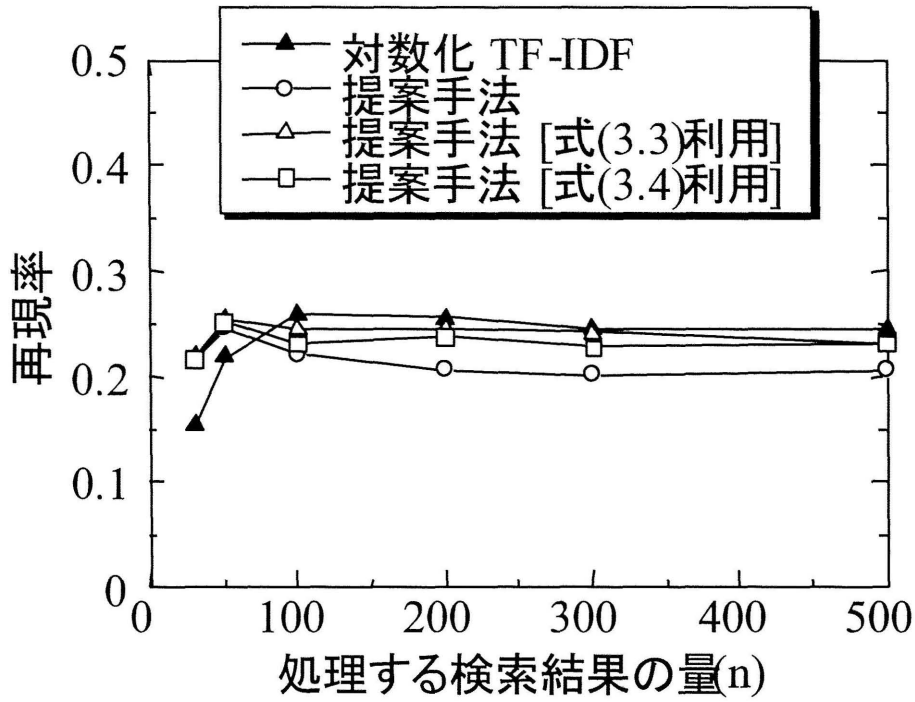

図 3.9 検索結果処理量と再現率の関係 (式 (3.3) および式 (3.4) を利用した場合))

3.5 応用事例

本節では，提案手法を実際のポータルサービスに応用した例を示すことで，提案手法の有効性を示す．

3.5.1 ニュース記事検索システム

提案手法をニュース記事検索システムに利用した例 [91] を図 3.10 に示す．この図では，「安倍晋三」というキーワードで検索を行った場合の画面例を示している．本システムでは，ニュースサイトから提供されるニュース記事を，提案手法を利用した検索システムに登録し，ニュース記事の検索サービスを提供している．

本システムは，ユーザから入力された検索キーワードに応じて，画面右側にニュース記事の検索結果を提示すると同時に，左側に提案手法で検索結果を解析し生成した検索結果のインデクスを提示する．これにより，ユーザは検索結果中に存在するトピックを概観できるとともに，所望の情報を簡単に絞り込む事ができる．実際にユーザが左のインデクス中のラベルを選択した場合には，「最初に入力された検索キーワード」と「選択されたラ

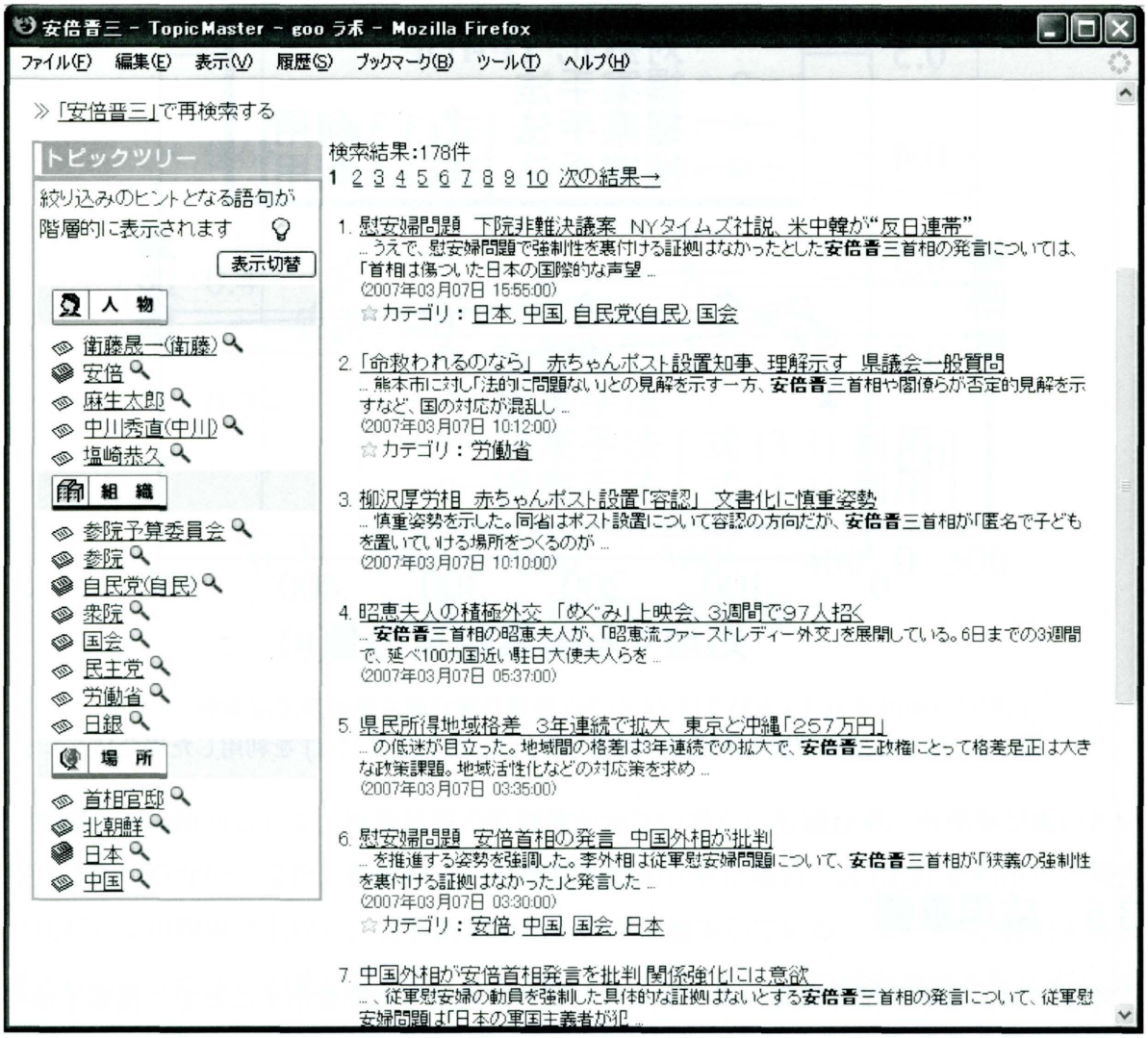

図 3.10 ニュース記事検索システムへの適用例 (検索条件「安倍晋三」)

ベル」の AND 条件で検索が行われ，右側に結果が提示されると同時に，その結果を元に生成されたインデクスが選択されたラベルの下部分に提示し，繰り返し絞り込みを行うことが可能となっている．図 3.11 に，「地震」というキーワードで検索を行い，「能登半島」というラベルを選択した場合の状態を示す．「能登半島」の下に提示されているラベルは，「地震」AND「能登半島」という検索結果により提示されたものであるため，より詳細な情報の絞り込みが可能なラベルが提示されていることがわかる．

また，図 3.12，3.13 に，Zeng らの手法 [70] および Vivisimo が運営する検索エンジン [87] と，検索結果のラベルを比較した例をあげる．これらは英語をベースとしたシステムであり，英語のニュース記事を検索対象としているので，検索条件を英語に訳して検索を

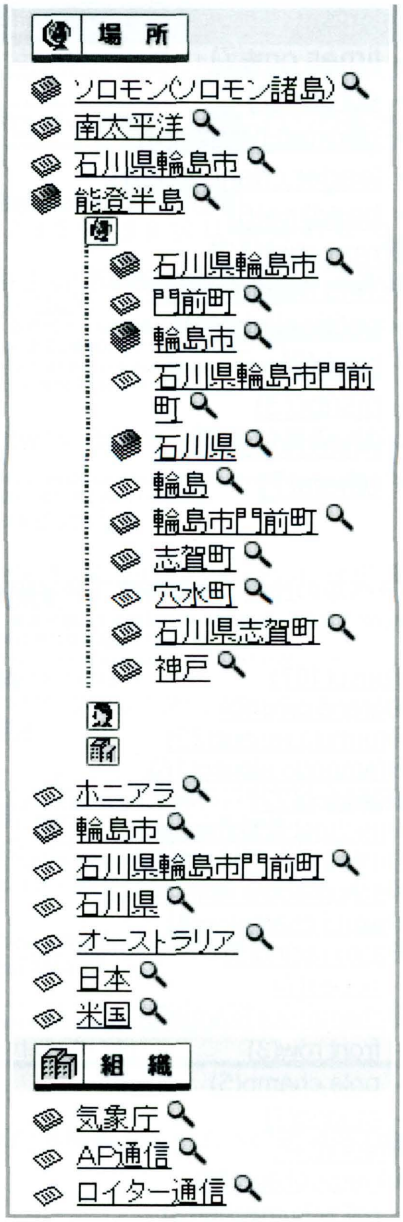

図 3.11 ニュース記事検索システムへの適用例 (検索条件「地震」)

行った結果を示している．結果を見ると，従来手法のラベルには，入力した検索キーワードと関係のない一般的なキーワードが含まれていたり，一般語と固有名詞が入り交ざっていたりで，検索結果を概観する事が困難である．それと比較して，提案手法では，明確なラベルが種類ごとに整理されており，概要の把握と言う点でより有益であり，絞り込み検索の候補を見つけるためにも有益であると考えられる．

図 3.12　ラベルの比較 (検索条件「ドイツ 総選挙」)

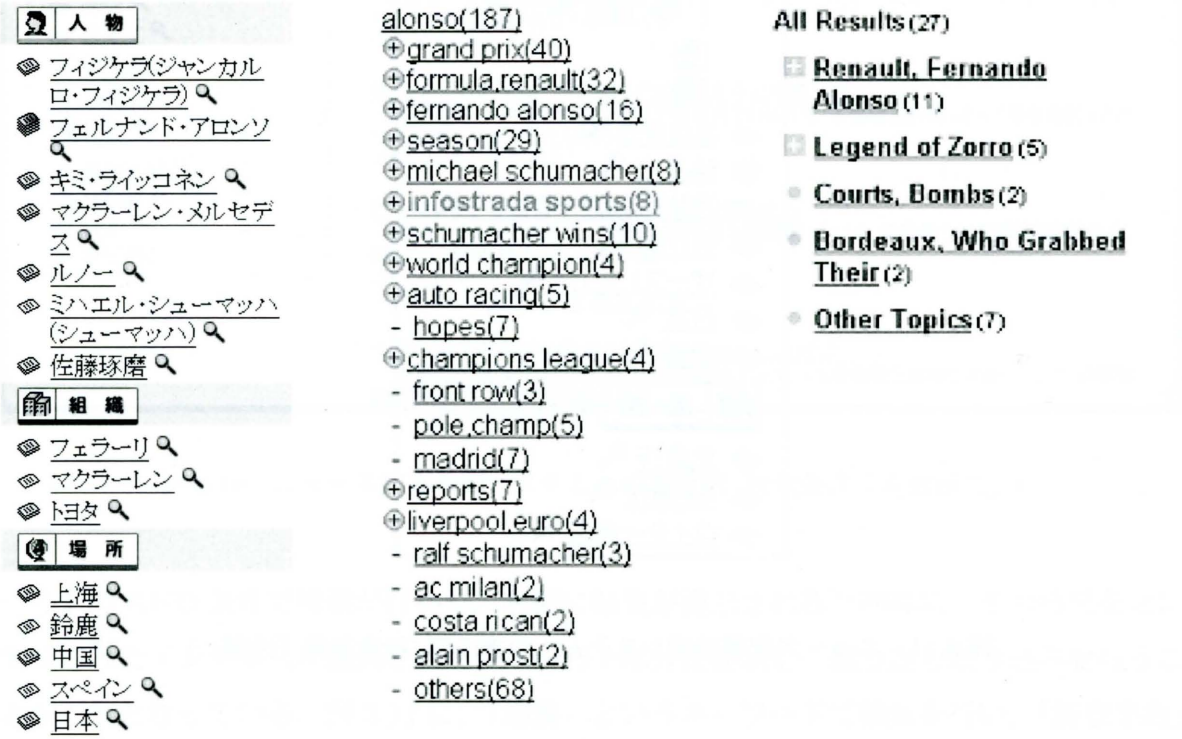

図 3.13　ラベルの比較 (検索条件「アロンソ」)

3.5.2　ブログ記事検索システム

ブログ記事検索システムに利用した例 [86] を図 3.14 に示す．この図では，「Wii」という検索キーワードで検索を行った結果を示す．このシステムでは，Web 中のブログサイ

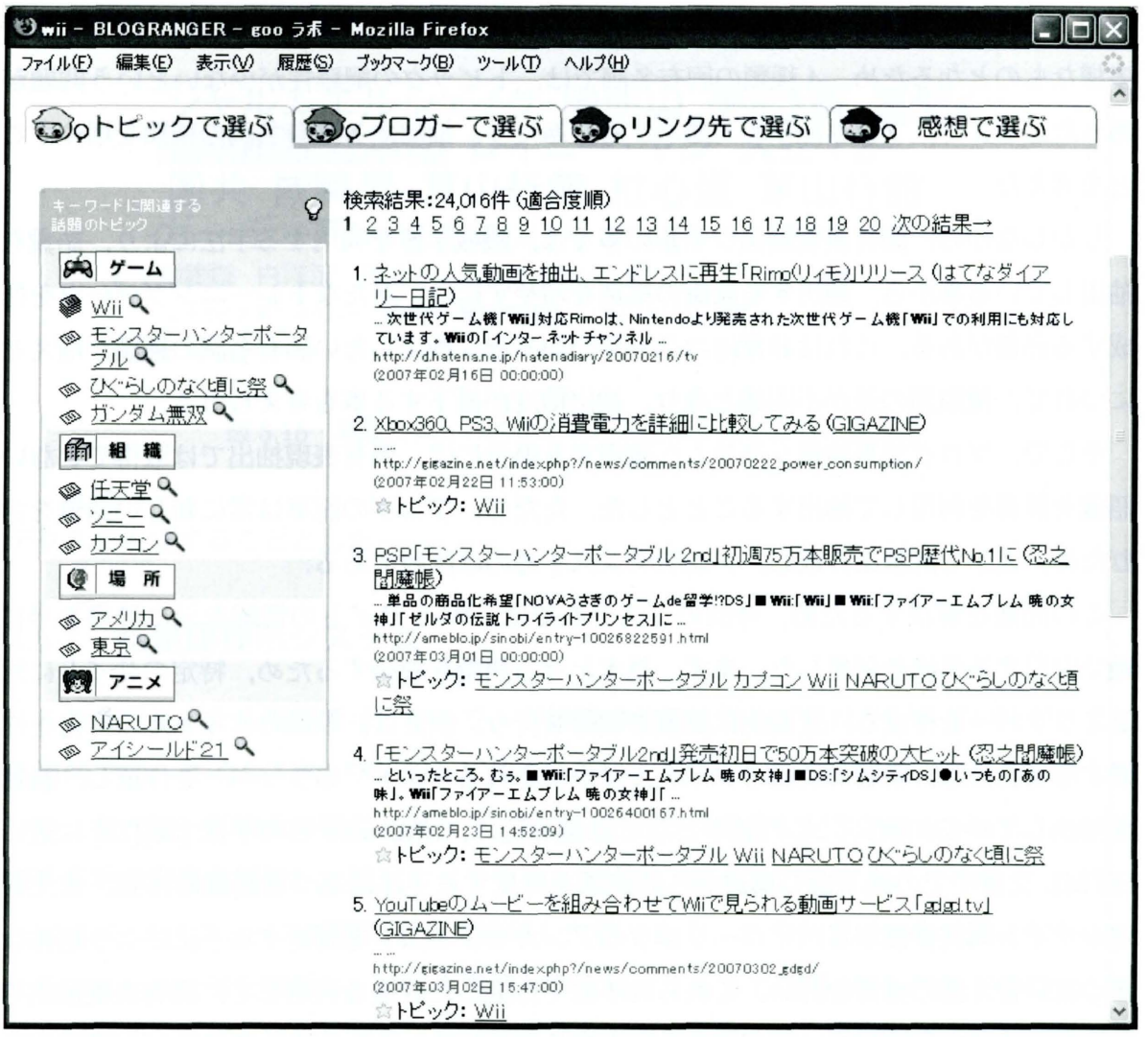

図 3.14 ブログ記事検索システムへの適用例 (検索条件「Wii」)

トをクロールし，ブログの記事単位にタイトルや本文を抽出した結果を検索システム登録している．基本的な操作はニュース記事検索システムと同様である．

ブログ検索システムに本手法を適用する上で，考慮しなければならなかった点は，以下の 2 点である．

- 抽出する語彙の種類
- 検索結果のランキング

まず，「抽出する語彙の種類」について述べると，ニュース記事を対象とした場合には，IREX[79] で規定された固有名詞の種類 (人名，組織名，地名，その他の固有物) である程

度のトピックをカバーする事ができたが，ブログ記事を対象とした場合には，トピックが多様なものとなるため，4種類の固有名詞では，トピックの網羅性が少ないという問題があった．そこで，ブログ記事を対象とするにあたり，抽出する固有名詞の種類を増やすことを考えた．

しかしながら，固有表現抽出ツールの多くは，機械学習を利用する手法により，語彙を抽出している事から，抽出する語彙の種別を増やすには，新たなトレーニングデータを作成する必要がある．これは非常にコストがかかる上，抽出したい固有名詞の種類が増えるにつれて，種類間の差が不明確となり，抽出精度が低下する事も考えられる．

そこで，ブログ記事検索システムへ適用する場合には，固有表現抽出では取得できない語彙を辞書を利用して抽出することとした．ただし，ブログの記事は常に新しい語彙を含むため，人手で辞書を作成し，メンテナンスするのは困難である．

この問題を解決するため，今回のシステムでは，Webサイトの情報を元に辞書を半自動で取得する手法を利用した．まず，基本となる語彙を取得するため，特定のサイトに対してラッパーを作成し，定期的に語彙の取得を行う．例えば，映画のタイトルの辞書を作成する場合には，新着の映画のタイトルが提供されるサイト[*3]のラッパーを作成し，語彙を抽出している．次に，この基本となる語彙を元に，Shinzatoらの手法 [56][78] に近いHTML文書中での繰り返し構造から同位語を発見する手法に基づき語彙の拡張，また特定のサイトへ共参照するハイパーリンクのアンカーテキストを解析する手法により略称などの言い替え語の獲得を行い，これらの手法で作成した辞書を利用して，固有表現抽出では取得できない語彙をブログ記事から抽出する事とした．本システムでは，ブログ記事中にはエンターテイメント系の情報が多いと考え，「テレビ番組」，「映画」，「ゲーム」，「本」，「アニメ」等の辞書を構築し，利用している．

次に，二点目の「検索結果のランキング」について示す．ニュース記事を対象とした場合には，いわゆるTF-IDF[53]でのランキングや，記事のタイムスタンプでランキングする事で，検索キーワードと関連する文書が検索され，ラベルの抽出も可能であった．しかし，ブログ記事の場合，記事の質は玉石混淆で，上記で示したランキングを用いた場合，しばしば検索条件に適合しない記事が多く検索され，有益なラベルの抽出ができない状態となった．そこで，ブログ記事検索に適用する場合には，Fujimuraらの提案するEigenRumorアルゴリズム [12] によって，人から参照されるような有益な内容を含む記事が上位になるような検索結果ランキングを行い，ランキング上位の検索結果を元にイン

[*3] http://movie.goo.ne.jp/schedule/upcoming.html 等

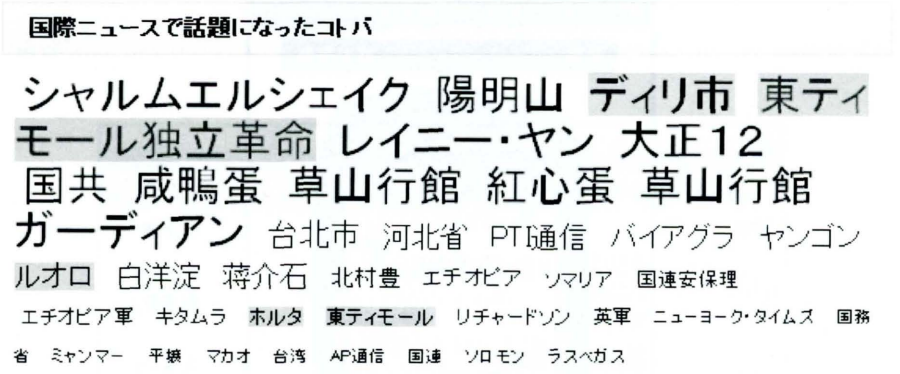

図 3.15 話題のニュースキーワード提示システムへの適用例

デクスを生成することとした.

3.5.3 話題語提示システム

これまで挙げた応用例では，キーワード検索の結果に対して処理を行っていたが，この応用例では，タイムスタンプに対して検索を行い，最新の記事を取得し，それを解析している．キーワード検索の場合と異なり，幅広いトピックを含む記事が検索されるため，比較的多数の記事を解析し，提示すべきラベルを決定，提示している．

図 3.15 は，ニュース記事に対して適用した例である．この例では約半日間に配信されたのニュース記事を解析し，ラベル選択基準を利用する事で話題のキーワードを抽出し提示している．ラベル選択基準の「検索結果の中での重要性」は，今回のような利用法でも当り前過ぎず，希少過ぎないワードを高く評価するために利用可能であり，「検索条件との関連性」は，今回の場合，過去にあまり話題になっていない語を高く評価する指標として有益である．このことから話題提示システムにおいても 3.3 で提案した基準をそのまま利用している．また，この応用例では，マウスカーソルをキーワードにのせた場合には，類似した文書集合で出現するキーワードをハイライト表示しており，より簡単にトピックの内容を知ることを可能としている．また，図 3.16 および図 3.17 は，ブログ記事に対して適用した結果であり，ブログ中でのトピックをランキング形式で提供し，興味のあるキーワードを選択するとそのキーワードに関連するブログ記事を簡単に閲覧できるアプリケーションとなっている．

このようなアプリケーションのメリットは，ユーザにとっては入力の手間なく話題の情報の概要が把握でき，キーワードを選択するだけで，より詳細な情報を検索する事ができる点である．また，サービス運営側にとっては，これらのサービス自身やこれらのサービ

図 3.16 話題のブログキーワード提示システムへの適用例

スから検索が行われる事によってアクセス数が増加する点がメリットとしてあげられる.

3.6 まとめ

本研究では，キーワードベースの検索システムにおいて，膨大な検索結果中に所望の情報が埋もれてしまうという問題に対して，検索結果中の主要なトピックを示すラベルを抽出し，検索結果とともに提示することで，検索結果をクラスタリングし，ユーザの検索を支援する手法について示した.

提案手法では，ラベルの候補となるタームの抽出に固有表現抽出を利用し，固有名詞の抽出を行う事を提案した．これは文書のトピックを特徴付ける情報として固有名詞が有益であると考えたためである．また，固有表現抽出によって抽出されたタームに付与される種類 (人名，地名等) をラベルの提示時に利用する事で，同種のラベルをまとめ, 構造化した形で提示できる．二つ目の提案は，上記で抽出したラベル候補からユーザに提示するラベルを選択する新たなラベル選択基準である．この基準では，検索結果内でのタームの重

図 3.17 話題のブログキーワード提示システム (モバイル版) への適用例

要性と，検索条件との関係性に基づき各タームを評価することで，絞り込み検索に有益であると考えられるラベルの抽出を可能とした．

提案手法の評価として，IREX のテストコレクションを用いた評価を行った．その結果，提案手法は，比較対象の手法より，検索結果を処理する文書量が少ない場合にも，絞り込み検索に有益なラベルを抽出できる事がわかった．これは，検索結果クラスタリングの計算コストが低下するだけでなく，検索システムから取得する検索結果の量も減少させることができるという点で，Web 上の検索システムなど，多くのユーザにアクセスされるシステムを考えた場合には非常に有利な特性である．

また，提案手法では，比較対象の手法と比較して全体的に高い適合率を示す事がわかった．また，提案手法の二つ目の仮説「検索条件との関連性が高いタームが重要」を表現する式を使い分けることにより，適合率を重視した手法や，適合率と再現率のバランスを重視した手法を使い分け可能であり，適合率を重視した場合には比較対象手法と比較して 15 ポイント以上の高い適合率を示し，バランスを重視した場合にはより少ない検索結果処理量で，比較手法と同等の適合率および再現率を示した．

さらに，本手法が，実際のポータルサービスに応用された事例である「ニュース記事検索システム」，「ブログ記事検索システム」，「話題提示システム」を示すことで，本手法の

有効性を示した.

第 4 章

グラフ分析を利用したトピック構造マイニング

4.1 はじめに

キーワード入力型の検索システムを利用する目的として，少量の検索条件にマッチする文書を取得したいという要求の他に，検索結果の全体集合から，概要や特徴的な情報を抽出したいという要求が考えられる．このような場合のユーザの要求としては以下の 2 点が考えられる．

- 検索結果中の主要なトピックが知りたい
- 検索結果中の個々のトピックを示す文書にアクセスしたい

これらのユーザの要求を満たす手段として，検索結果等の文書集合から重要キーワードを抽出することで，文書集合中のトピックを提示する手法 [69][70][16][60][80] や，クラスタリングアルゴリズムを利用して，文書集合をクラスタ化する手法 [7][39][24] 等が挙げられる．

しかし，上記の手法でトピック抽出や文書のクラスタ化ができたとしても，以下の問題点が存在すると考えている．

問題 1 トピックへのアクセス時の問題

検索結果中に多くのトピック (クラスタ) が存在する場合，特定のトピックへのアクセスやトピック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

個々のトピック (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

ここで言うトピックとは，検索結果等の文書集合中の文書で表現された実世界のイベントもしくはアクティビティを示している．特に本章では，特別な説明がない限り，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トピックと呼ぶ．また，本章において，検索結果は個々の結果を個々の文書と見なした文書集合と見なす．

以上の問題意識から，本研究では，文書集合中に含まれるトピックへのアクセスおよび特定のトピックに関連する文書へのアクセスを支援するトピック構造マイニング手法を提案する．

提案手法では，まず文書集合から以下に示すノードとエッジで構成されるグラフ構造(文書集合グラフと呼ぶ)を構築する．

ノード 文書集合中の各文書

エッジ 文書間の関係[*1]

次に，構築したグラフ構造の各ノードに対して，ノード間の結合度合を元にノードの重要性を評価する中心性を算出する．そして，各ノードの中心性，ノード間のつながりから，各ノード(文書)を以下の4タイプに分類する[*2]．

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持つノード．隣接ノード群が示すトピックの中心的な内容を最も良く表す文書を示す．

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．コアノードが示すトピックの中心的な内容を補足する文書を示す．

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．トピックの中心的な内容と関連性はあるが，他文書とは異なる情報を表す文書を示す．

アウトライヤーノード 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を示す．

図4.1に文書集合グラフおよび，上記に示す4タイプのノードの概念図を示す．図中のノード“ax”(“bx”)は，トピック“a”(“b”)に関するノードである事を示す．図にあるよ

[*1] エッジのある/なしおよびエッジの重みは，文書間の類似度を元に決定する．

[*2] 本提案手法では，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．

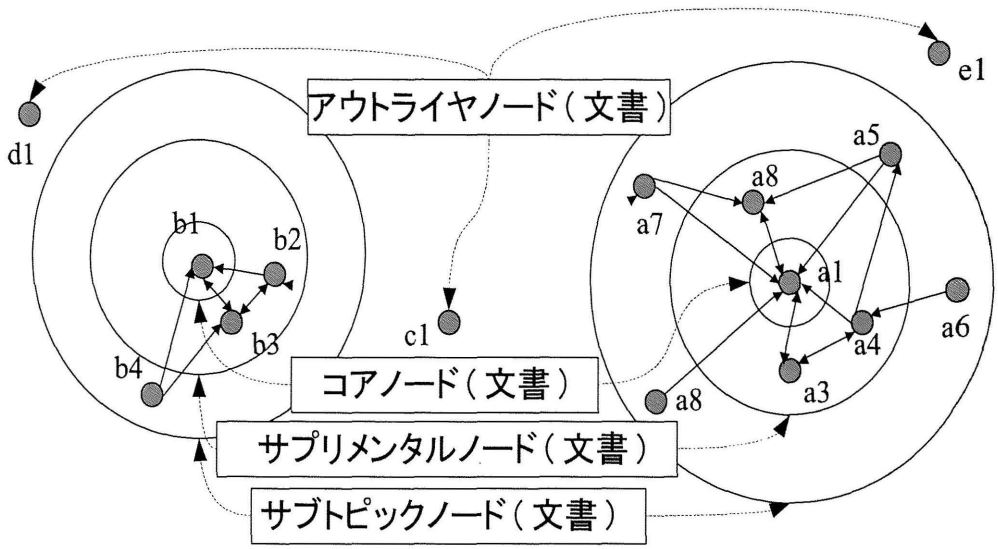

図 4.1 文書集合グラフの概念図

うに，グラフ構造中で同一のトピックに関連するノードは，そのトピックを示すコアノードの回りに階層的に配置される．

以上に示した「文書集合グラフ」「各ノードの中心性」「各ノードのタイプ」を利用することで，前述した問題に対して以下のような解決策を提供する．

問題 1 に対しては，文書集合内のコア文書を，中心性のスコア順に提示する事で，文書集合中の主要なトピックを提示し，トピックの閲覧を支援する．また，文書集合グラフと各ノードの中心性を用い，文書集合の関係を可視化する事で，文書やトピックのつながりを視覚的に閲覧する事が可能となり，トピック間のつながりの強さ，つながりの内容を知ることができる．

また，問題 2 に対しては，文書集合グラフと各ノードの中心性を同時に用いて，コア文書に関連する文書を集めることで，特定のトピックに関する文書クラスタを生成する．加えて，クラスタメンバとなる個々の文書には，上述の文書タイプが付与されているため，“トピックの中心的な内容を最も良く示す文書” や “トピック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” 等を選択的に閲覧することができる．

評価は，2 つの観点から行う．1 点目は，提案手法の基本特性評価であり，提案手法をトピック抽出，クラスタリングに適用した場合の精度を評価する．2 点目は，トピック構造マイニングの基本となる文書のタイプ分類が意図通りに機能しているかを評価する．また，グラフ構造とノードの中心性を利用した文書集合の可視化結果を示し，提案手法で考

えるトピック構造が「抽出したトピック間の関係の明確化」等，文書集合のブラウジングやマイニングに有益である事を示す．

さらに，近年ニュース記事やブログ記事のようにタイムスタンプ付きの文書が多く普及しており，このような文書を処理するには，タイムスタンプを考慮した分析法が必要であると考え，上記に示した「文書集合グラフ」構築時に，文書間のタイムスタンプの近さを考慮する手法についても検討を行った．

以下，4.2 で，提案手法の詳細について示し，4.3 で評価について述べる．4.4 では可視化の有効性について示す，4.5 では，提案手法をタイムスタンプ付きの文書へ応用した手法について示し，4.6 でまとめる．

4.2 提案手法

ここでは，検索結果等の文書集合中のトピックや文書に効率的にアクセスするためのトピック構造マイニング手法について示す．

まず，提案手法では，文書集合から以下に示すノードとエッジから構成される重み付き有向グラフ (文書集合グラフと呼ぶ) を構築する．

ノード 文書集合中の各文書
エッジ 文書間のつながり

次に，この文書集合グラフ中の各ノードについて，ノード間の結合度合を元にノードの重要性を評価する中心性を算出する．そして，中心性のスコアと，文書集合グラフにおけるノード間のつながりの関係を利用する事で，各ノードを 4 つのタイプに分類する．

以下，4.2.1 では文書集合グラフの構築法について示す．4.2.2 では，グラフ構造のノードの中心性スコアの計算方法について示す．また，4.2.3 では，上記で作成したデータに基づくノードの意味付けと，それらを用いた提案手法の利点について示す．

4.2.1 文書集合グラフ構造の構築

文書集合グラフは，文書集合中の文書間の関係を示すグラフである．この構築にあたっては，密に結合したノード間の関係を選択的に抽出し，いわゆる「その他」にあたる文書の影響を無視したいとの考えから，高い関連性を持つ関係のみを抽出し，グラフ構造を構成する方法が必要である．そこで，本研究におけるグラフ構造の構築では，Kamvar によって提案された “Interested Reader Model”[27] をベースとして利用する．これは

PageRank[4] の "Random Surfer Model" に類似したモデルであり，いくつかのトピックに関する文書で構成される文書集合内の文書を次々に読み進める Reader を仮定したモデルである．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．

$$N = (A + d_{max}E - D)/d_{max} \tag{4.1}$$

ここで，E は単位行列を示し，D は対角行列であり，以下の式で計算される．

$$D_{i,i} = \sum_j A_{i,j} \tag{4.2}$$

また，d_{max} は D の要素のうち最大の値を取るものである．A はノード間の類似度を表現する隣接行列であり，以下の式で定義される．

$$A_{i,j} = \begin{cases} sim(i,j) & if \quad j \in TopSim_p(i) \\ 0 & otherwise \end{cases} \tag{4.3}$$

ここで，$TopSim_p(i)$ は文書 i との類似度が高い文書 p 件に含まれる文書の集合を示す．$sim(i,j)$ は，文書を対数化 TF-IDF 重み [76] によるキーワードベクトルとして表現した場合の文書 i と文書 j のコサイン類似度を示している．

式 (4.1) の最初の項 (A) は，上記 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．

しかし，以上の仮定に基づいたグラフ構造では，すべてのノードからの外向きのエッジが同じ本数 (p 本) となるため，しばしば文書間の類似度が非常に小さいにもかかわれずエッジが存在する事がある．提案手法において，エッジは高い関連性を持つ関係を示すためのものであるため，類似度が低い文書間にエッジが張られる事は本来の目的に反する．そこで，外向きエッジのうちごく少ない重みしか持たないエッジを除去する事を考える．この操作を以下の式で示す．

$${N'}_{i,j} = \begin{cases} N_{i,j}/l_{i,q} & if \quad j \in TopLink_q(i) \\ 0 & otherwise \end{cases} \tag{4.4}$$

$l_{i,q}$ はノード i からの外向きエッジを遷移確率の降順に並べ，閾値 q を越えるまで加算した遷移確率の合計値を示し，$TopLink_q(i)$ は，その時，加算対象になったエッジの接続先

ノードの集合を示す．$q = 1$ の場合には，$N' = N$ である．また，以後 q を不要エッジ除去係数と呼ぶ．

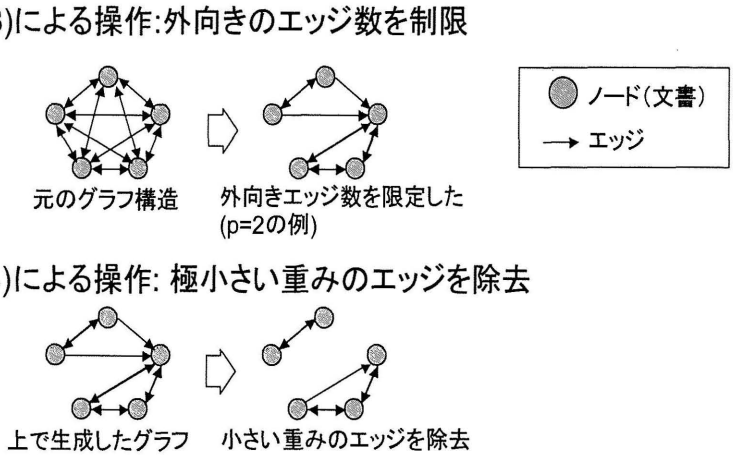

図 4.2 不要エッジ除去処理

図 4.2 に，上式 (4.3) および (4.4) で行われる不要エッジ除去処理をまとめる．この処理を行うことで，2 つのノード間は以下のいずれかの関係で示される．

- 双方向エッジが存在
- 片方向エッジのみが存在
- エッジなし

4.2.2 中心性スコアの算出法

本節では，前節で作成した文書集合グラフを利用して，文書集合中に含まれるトピックの中心的な内容を含む文書を特定するために，グラフ構造の中心性を算出する方法を示す．このような文書を特定する事で，文書集合中のトピックや，そのトピックに関連する文書の抽出を可能とする．

そもそも本研究における「トピック」とは，複数の文書で表現されたイベントやアクティビティを指しており，「トピックの中心的な内容」とは，複数の文書で共有される内容である．このことから，「トピックの中心的な内容」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言える．これを特定のトピックに関する文書群に特化して考えると，最も「トピックの中心的な内容」を保持する文書は，その文書群内のより多くの文書とより強い類似度を持つ文書であるといえる．これを前節で構成したグラフ

構造上で考えると，特定のエリア内でより多くの強いエッジを持ったノードに該当する．また，単にエッジを多く持つだけではなく，よりトピックの中心に近い文書 (ノード) と強くつながっていることも重要な要素となると考えられる．

以上の考えから，提案手法では，入次数，出次数，closeness[46]，HITS[31]，PageRank[4] 等，様々に定義される中心性のうち，PageRank を利用する．以下に，PageRank の算出式を示す．

$$S(V_i) = (1-d) \times \sum_{V_j \in IN(V_i)} \left(\frac{1}{|OUT(V_j)|} \times S(V_j) \right) + d \tag{4.5}$$

ここで，$S(V_i)$ はノード V_i の中心性スコアである．$IN(V_i)$ は V_i に対してエッジを張っているノードの集合であり，$OUT(V_i)$ は V_i からエッジを張られているノードの集合である．d はランダムジャンプの確率を設定するダンピングファクターである．このランダムジャンプは，通常ノード間のつながりに基づき遷移する Walker が，一定の確率でつながりを無視し，ランダムに次のノードに遷移する動きをモデルに導入している．

本提案手法においても，ランダムジャンプを利用した PageRank を用いてノードの中心性スコアの算出を行う．PageRank と異なるのは，エッジ重みが均一ではない点であり，中心性スコアの計算式は以下の式で算出される．

$$S(V_i) = (1-d) \times \sum_{\forall j} (N'_{j,i} \times S(V_j)) + d \tag{4.6}$$

4.2.3 文書集合グラフと中心性スコアを利用したトピック構造マイニング

ここでは，以上の手法により構築された文書集合グラフと，そのグラフ中のノードの中心性から導かれる各ノードの意味とその抽出法，および以上から構成されるトピック構造を目的の達成に利用する方法について示す．

本研究で提案するマイニング手法では，XY 平面に，文書集合グラフを配置し，Z 軸上にノードの中心性スコアを割り当てた 3 次元のグラフ構造 (「3D 文書集合グラフ」と呼ぶ) を考える．例を図 4.3 に示す．ここで，ノード “ax”(“bx”) は，トピック “a”(“b”) に関するノードである事を示す．また，XY 平面へグラフの配置では，エッジでより強く繋がるノード同士を近くに配置し，繋がりのないノード同士は遠くに配置する．

文献 [33] で示される手法は，中心性のスコアを単純に利用し，文書のランキングを行っている．一方，本研究の目的を達成するには，トピック毎に文書を分類する事が必要となる．しかし，中心性のスコアのみでは要素を分類することが出来ない．つまり，図 4.3 中のノードを中心性のスコアによって順位付けすると，“a1, a2, a3, b1, b2, a4, ...” という順序が取得され，トピック “a” と “b” のノードが混在する．そこで本研究では，トピック

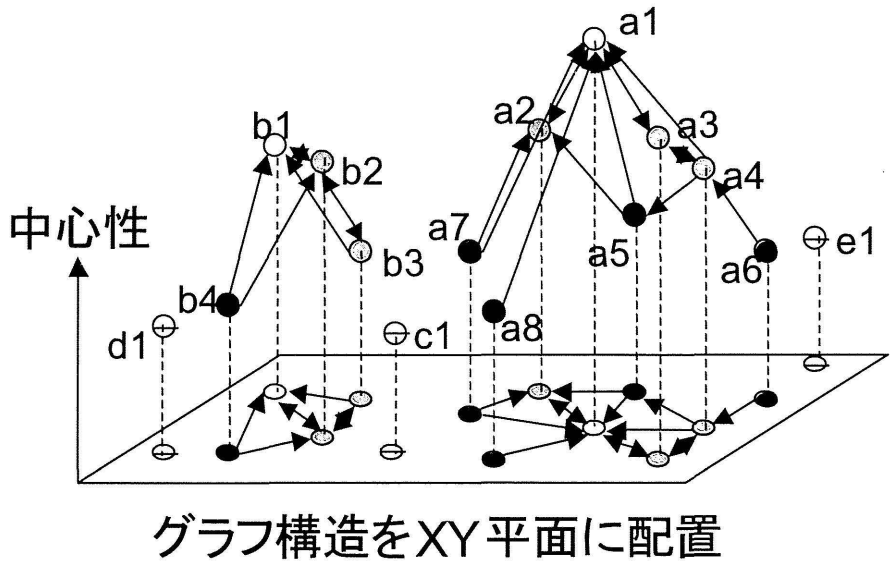

図 4.3 文書集合グラフとノードの中心性を利用した 3D 文書構造グラフの例

に応じた文書の分類を行うために，文書集合グラフの構造とノードの中心性の両方によって生成された山状の構造を利用して，トピック構造の抽出を行う．

ここで，グラフ構造と中心性スコアの関係について考える．中心性スコアの定義によれば，多くのエッジが存在するエリアのノードは高いスコアを持つ．そ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トピックに関連するノードで構成される．したがって，図 4.3 のそれぞれの山は，それぞれ異なるトピックに対応すると考えられる．またこの山に含まれるノードの位置に応じて，文書にはそれぞれ特徴があると考えられる．以下では，それぞれのノードをグラフ構造中の位置から 4 タイプに分けて，それぞれのノードに該当する文書の特徴を説明する．

最初のノードは図 4.3 で，山の頂上にあるノード (a1 や b1 に該当) である．このタイプのノードは，周囲のノードと比較して中心性スコアが最も高いノードであることから，周囲のノードから最も高い状態遷移があり，周囲のノードと最も関連性の高いノードであると言える．この点からこのノードが示す文書は，そのエリアのトピックの中心的な内容を最も良く表現する文書である．本研究ではこの文書 (ノード) をコア文書 (ノード) と呼ぶ．4.3.3 の実験では，中心性スコアとトピックの中心的な内容の網羅性の関係について評価を行い，コア文書が実際にトピックの中心的な内容をよく含む事を検証している．

2 つ目のノードは，頂点と近接したノード (図 4.3 中の a2, a3, a4 や b2, b3) であり，

コアノードから双方向エッジのみをたどって到達できるノードである．双方向エッジは，相互にリンクが張られており，高い関連性を示す．これらのノードはコアノードとの間で高い確率で状態遷移があり，文書の内容もコア文書と高い関連性があると考えられる．本研究ではこの文書 (ノード) をサプリメンタル文書 (サプリメンタルノード) と呼ぶ．ただし，直接コアノードとつながっているノードと，間接的にしかコアノードつながっていないノードでは，それぞれのノードにより表現される文書が持つ特性に違いがある事も考えられるため，4.3.3 にて，その違いについて検証した．

3 番目のノードは，図 4.3 の a5, a6, a7, a8, b4 のようにコアノードもしくはサプリメンタルノードにエッジを張るノードである．外部のノードへの状態遷移や自己遷移と比べて，特定のトピックのコアノードやサプリメンタルノードへの遷移確率が高いノードがこの種のノードに該当し，必ずしもトピックの中心ではないがトピックに関連する情報を含んでおり，トピックの周辺の情報等ノベルティの高い情報を含む事が多いノードであると考えられる．本研究ではこの文書 (ノード) をサブトピック文書 (サブトピックノード) と呼ぶ．4.3.3 の実験では，サブトピックノードを選択することで，上記に示す通りノベルティの高い情報 (関連文書内で希少性が高い情報) を含む文書を選択できるか検証した．

最後のノードは，どのトピックのノードに対しても強い関連性がないノードである．図 4.3 の c1,d1,e1 がこれにあたる．このノードは，他に似ているノードが少なく，自己遷移確率が高い．本研究ではこの文書 (ノード) をアウトライヤー文書 (アウトライヤーノード) と呼ぶ．

以上で示したノードのうちグラフ構造中で特定のトピックに関連するノードは，該当するトピックを示すコアノードの回りに階層的に配置され，アウトライヤーノードは，その階層の外に配置される (4.1 の図 4.1 参照)．また，表 4.1 に，上記で示したノードの特徴および，ノードが表現する文書の特徴をまとめる．個々のノードタイプ名の左に示したイメージは，図 4.3 中で該当するノードを示す．

以上で示した構造を利用することで，ユーザは選択的に所望のトピック，文書へアクセスする事が可能となる．例えば，コアノードのみを選択することで，文書集合中の主なトピックをピックアップしたり，特定のトピックの中心的な内容を最もよく表現する文書を選択的に閲覧できる．また，サブトピック文書を閲覧することで，予期しないノベルティの高い情報を提供する文書を閲覧すること等も可能である．

また，山状に配置されたノード群をクラスタと見なすことで，クラスタリングと同様の効果を得ることもできる．その場合には，コア文書やそのタイトルを個々のクラスタを表

表 4.1 ノードの特徴およびノードが表現する文書の特徴

ノードタイプ	グラフ構造中でのノードの特徴	ノードが表現する文書の特徴
○コアノード	頂点ノードであり、周辺ノードとの結びつきが最も大きい	それぞれのクラスタを最も良く表現する
◎サプリメンタルノード	双方向エッジで、頂点ノードと接続する	コア文書と類似し、それを保管する内容を含む
●サブトピックノード	コアノードおよびサプリメンタルノードに対して高い結びつきがある	コア文書と関連があり、かつなんらかの新規性ある内容を含む
⊖アウトライヤノード	自身以外のノードと強いつながりがない	独立した内容を示す

現するラベルとして利用できる．また，特徴的なキーワードをラベルとして利用する場合は，単純にクラスタ中の文書が最も多く含むキーワードを選択する事で，ある程度の精度でキーワードを抽出する事が可能である．表 4.2 には，新聞コーパスに対して「テロ or 爆破 or 爆弾」で検索を行った結果上位 200 件を利用して，提案手法でクラスタを作成した場合に抽出されたラベルの候補を示している[*3]．コアノードの中心性が高いものから 5 つのクラスタの例である．キーワードと固有表現は，各クラスタ中での文書頻度の高い形態素 (形態素解析器 茶筅[*4]により自立名詞もしくは未知語と判定された語)，固有表現 (磯崎らの手法 [26] のツールを利用して出力された語) をそれぞれ，最大 10 個と 5 個取得した．詳細な評価ではないが，これらの例から，どのラベル候補を利用したとしても，ある程度クラスタの内容を類推する事が可能である．これは，従来のクラスタリング手法の場合には，適正なクラスタ数や類似度の閾値を指定する事が困難であり，複数のトピックを含むクラスタや不自然に分割されたクラスタが出力され，その結果，クラスタを説明するラベルの意味も不明瞭となるのに対して，提案手法はトピックが集まっている部分を選択的に抽出する手法であるため，抽出されたクラスタはある程度同じトピックに関する文書となっているためであると考えられる．

さらに，上記の山状の構造を利用する事で，2 つのトピック (2 つの山) の関係についても知ることができる．例えば，サプリメンタル文書やサブトピック文書で山同士が連結し

*3 94，95 年の毎日新聞記事に対してキーワード検索した結果をクラスタリングした例．コーパス全体 (約 20 万文書) で文書頻度が 5,000 以上だった形態素，固有表現は一般語としてあらかじめ除去している．また，検索条件とした語も表 4.2 からは除去している．

*4 http://chasen.naist.jp/hiki/ChaSen

表 4.2 ラベルの例

クラスタ ID	情報タイプ	結果
1	コア文書タイトル	アルジェで爆発、10 人が死亡
	頻出キーワード	アルジェ, 負傷, 自動車, アルジェリア, イスラム, 犯行, 首都, 原理, 過激, 集団
	頻出固有名詞	アルジェ市, アルジェ, アルジェリア
2	コア文書タイトル	死者、約 200 人に –米連邦政府ビル爆破テロ事件
	頻出キーワード	オクラホマ, シティー, 連邦, ビル, 救出, 現場, ニューヨーク, FBI, 生存, がれき
	頻出固有名詞	オクラホマ, ニューヨーク, FBI, 連邦捜査局, キャロル・ビラレル
3	コア文書タイトル	バス爆破、6 人死亡 ハマスの犯行か –イスラエルの都市テルアビブ死者、約 200 人に
	頻出キーワード	パレスチナ, イスラエル , イスラム, バス, 西岸, ヨルダン川, ハマス, ガザ, 原理, 地区, 自治
	頻出固有名詞	イスラエル, パレスチナ, ハマス, ヨルダン川西岸, エルサレム
4	コア文書タイトル	死者 28 人に—ユダヤ人協会ビル爆破テロ事件
	頻出キーワード	ブエノスアイレス, ユダヤ, アルゼンチン, ビル, 当局, 負傷, メキシコ, 死者, イラン, 強力
	頻出固有名詞	ブエノスアイレス, アルゼンチン, ユダヤ, ユダヤ人協会, イラン
5	コア文書タイトル	パリの連続爆弾テロ事件で、イスラム武装団の 幹部を逮捕–フランス警察当局
	頻出キーワード	パリ, 武装, ミッシェル, アルジェリア, イスラム, 郊外, 過激, GIA, 原理, 高速
	頻出固有名詞	パリ, アルジェリア, GIA, サン・ミッシェル, パリ市

ている場合には，トピック間に高い関連性があると言える．逆に連結していない場合や，連結部分のノードがアウトライヤーであった場合にはトピック間の関係は薄いというような情報が得られる．また，その連結部分にある文書は二つのトピックの接点であるため，その文書を参照することで，つながりの意味を知る事ができる．

4.3 評価

本章では，2種類の評価を行う．1つ目は，提案手法の基本特性評価として提案手法をトピック抽出およびクラスタリングに適用した場合の精度を評価した．これにより，基本特性およびパラメータ操作に対する特性の変化についての知見を得る．2つ目は，トピック構造に関する仮説の検証として，4.2.3で提案した文書のタイプ別けが仮説通りに機能しているかについて評価した．全ての評価で新聞記事コーパスを利用して評価を行った．また，クラスタリングの評価では，NTCIR-4 WEB D(トピック分類タスク)[9]と同等の評価も行った．

4.3.1 評価リソース

まず，新聞記事を利用した評価のリソースについて述べる．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．

コーパスの各文書に対して著者がラベル付けを行った．それぞれのラベルは文書中で主に述べられているトピックを示す．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トピックのみを集めた主要トピックリストを作成した．テストセットおよび主要トピックリストの仕様を表4.3に示す．

一方，NTCIR-4 WEB D[9](以下単にNTCIRと呼ぶ)の評価では，NTCIRで提供されるNW100Gを利用した．実際には，11の検索課題についてNW100Gを検索した検索結果，それぞれ200件が提供されており，これらを利用してNTCIR-4 WEB Dと同じ評価を行い，タスク実施時に上位にランクされたシステム[47]との比較を行った．

表 4.3 評価に利用した新聞記事テストセットおよび主要トピックリストの仕様

テストセット名	検索語	文書数	ラベル数	ラベル付けされた文書数
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 4.4 実験条件

パラメータ	値
p: 1 ノードからの外向きエッジ数	2, 3, 4, 5, 6, 7
q: 不要エッジ除去係数	0.5, 0.6, 0.7, 0.8, 0.9, 1

4.3.2 基本特性の評価

新聞記事を利用したトピック抽出の評価

本節では，提案手法においてコア文書のみを抽出した場合に，そのコア文書群が表すトピックが，文書集合中に存在するトピックをどの程度を網羅しているか，また重複なくトピックを抽出できているかの点から評価を行った．

実験条件を表 4.4 に示す．実験は，p と q の全ての組合せで行った．p は一つのノードからの外向きエッジの数である．提案手法では，文書間のつながりを元にトピック間のつながりの取得を考えているため，一つの文書が複数のエッジを持つ事を前提とし $p = 2$ 以上で実験を行った．また，予備実験からある程度以上 p を大きくしても大幅な傾向の変化がなかったため，$p = 7$ までとした．また，q は 関連性の薄いエッジを除去するためのパラメータである．$q = 1$ の場合，不要エッジの除去は行なわれない．実験の結果，q をあまり小さくすると，基本的な特性が悪化したため，$q = 0.5$ までとした．

また，クラスタリングを利用した手法を比較対象とした．利用したクラスタリング手法は以下の二つである．

- K-means 法
- 凝集法 (セントロイド法)

クラスタリングを用いた実験では，正解セットから求めた理想的なクラスタ数[*5]でクラスタリングを行い，それぞれのクラスタで重心に最も近い文書を抽出し，その文書に付与されたラベルを抽出されたトピックとみなす．クラスタのノードが 2 つの文書で構成され，2 つの文書が異なるトピックのラベルを持っている場合には，2 つのうち正解のトピックが存在すればそれを選択することとした．

図 4.4 に，$q = 1$ の場合，つまり不要エッジの除去を行わない場合のトピック抽出結果を示す．グラフ中の全ての点は，10 個のテストセットによる評価結果の平均となっている．全体的な傾向として，適合率は高いが再現率が低いという傾向がある．

この再現率が低い理由としては，この条件では 1 つのノードからの外向きエッジの数が同じであるため，テストセット中で，ひとつのトピックに関連付く文書が少ない場合 (各クラスタが小さい場合) に，コア文書になるべき文書が異なるトピックの文書とエッジでつながり，コア文書として抽出できなかったのではないかと考えられる (この現象を「不要なエッジによるコアノードの埋没」現象とする)．実際，図 4.4 で示すように p が小さくなるにつれ再現率が上昇する傾向や，また一つのトピックを示す文書数が 3 未満である事が少ない「scandal」コーパスを処理した場合には $p = 2$ や 3 で比較的高い再現率 ($p = 2$ の場合 0.8182，$p = 3$ の場合 0.7727) を示す傾向が見られ，この理由を裏付けている．

また，図 4.5 下のグラフには，q の値を変化させ，関連性の薄いエッジを除去した場合

[*5] このクラスタ数は，それぞれ正解セット中の「主要トピック数 + 主要トピックに関係しない文書の数」で算出した値を用いた．

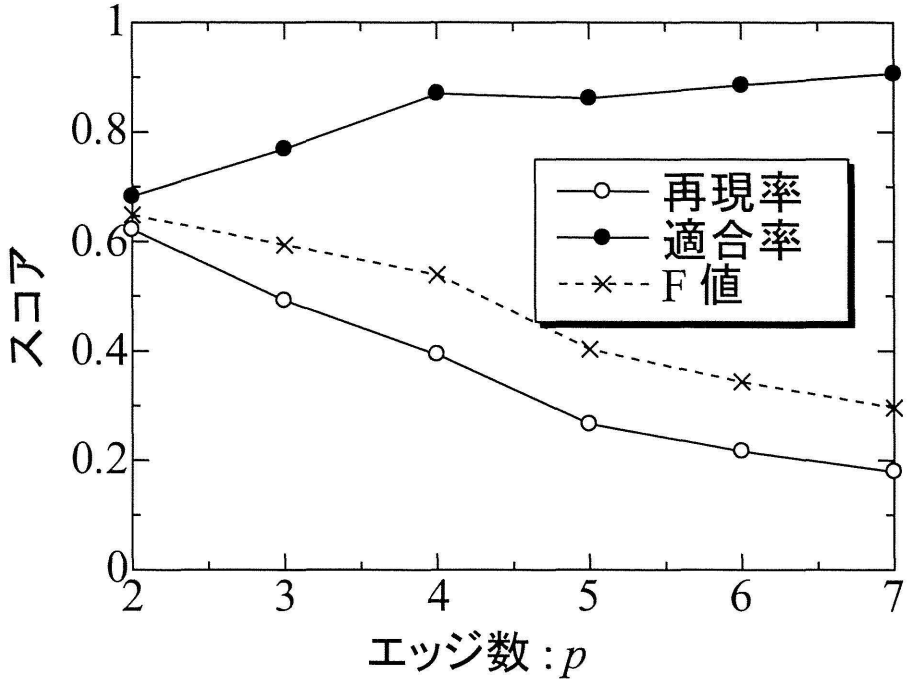

図 4.4 エッジ数 (p) とトピック抽出精度の関係 ($q = 1$ の場合)

のトピック抽出精度 (F 値) を示している．グラフ中の全ての点は，10 個のテストセットを利用して評価した結果の平均である．このグラフから，どのエッジ数 p を利用した場合にも，F 値は q を 1 から減少させると，$q = 0.7$ でピークを迎え，それ以下では，逆に精度が低下している事がわかる．また，ピークにおける F 値は p にかかわらずほぼ同じ値を示す．図 4.5 上に示す再現率も同様な傾向を示し，$q = 0.7$ でピークを迎え，それ以下では，低下している事がわかる．また，図 4.5 中に示す適合率は，q を下げると減少する傾向にある．ただし，再現率の上昇と比較するとその幅は小さいため，F 値は再現率と同様の傾向を示している．

p を変化させた場合の傾向を観察すると，p が小さい方が，再現率の変化は小さく，絶対値も高くなる．これは p が小さい場合には，前述の「不要なエッジによるコアノードの埋没」が起きにくいためだと考えられる．一方，適合率は，p が小さい場合に低くなる傾向がある．これはエッジ数が少ない場合には，一つのトピックを細かく分割し過ぎるためだと考えられる．以上の傾向は，q が大きい場合に顕著である．また，F 値に関しては，全体的な傾向として p が小さい方が値の変化が少ないが，最も高い F 値を記録する $q = 0.7$

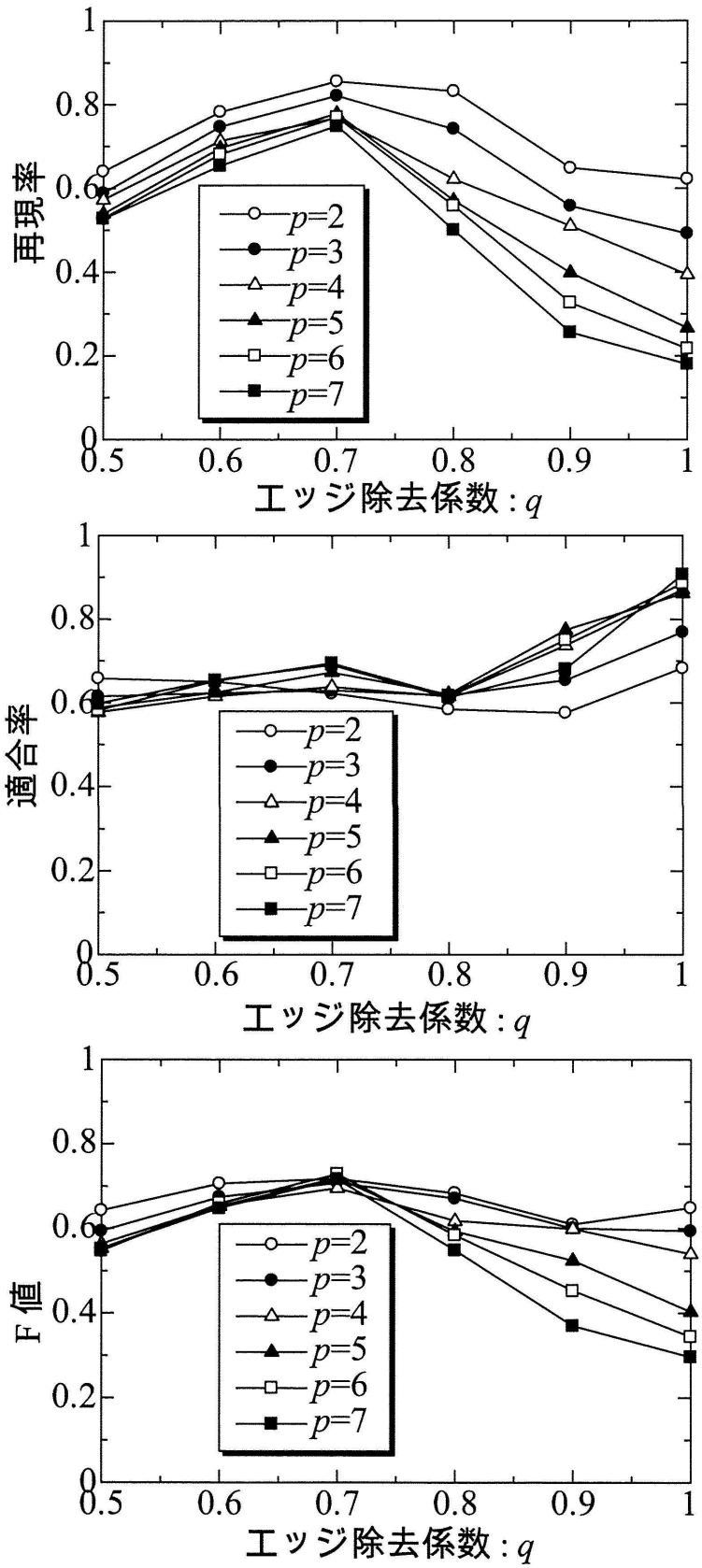

図 4.5 不要エッジ除去係数 (q) とトピック抽出精度の関係

表 4.5 トピック抽出精度の評価結果 (F 値)

手法	実験条件	平均	murder	terrorism	scandal	s+k
K-means	—	0.5839	0.5333	0.4783	0.6531	0.6167
凝集法	重心法	0.6597	0.6939	0.5116	0.6667	0.7272
提案手法	$p = 3, q = 0.5$	0.5937	0.6122	0.3750	0.6786	0.6607
提案手法	$p = 3, q = 0.6$	0.6737	0.7018	0.5614	0.6557	0.7107
提案手法	$p = 3, q = 0.7$	**0.7080**	**0.7333**	**0.600**	**0.6557**	**0.7442**
提案手法	$p = 3, q = 0.8$	0.6708	0.6885	0.5000	0.7843	0.7759
提案手法	$p = 3, q = 0.9$	0.6005	0.6087	0.5306	0.7727	0.6415
提案手法	$p = 3, q = 1$	*0.5936*	*0.5854*	*0.4000*	*0.7907*	*0.6947*
提案手法	$p = 5, q = 0.5$	0.5523	0.5106	0.4000	0.6786	0.6422
提案手法	$p = 5, q = 0.6$	0.6520	0.6909	0.4906	0.6780	0.6949
提案手法	$p = 5, q = 0.7$	**0.7194**	**0.6909**	**0.5818**	**0.7925**	**0.8033**
提案手法	$p = 5, q = 0.8$	0.5933	0.6000	0.5098	0.7556	0.6923
提案手法	$p = 5, q = 0.9$	0.5237	0.5128	0.4324	0.6471	0.5747
提案手法	$p = 5, q = 1$	*0.4032*	*0.2500*	*0.2941*	*0.6250*	*0.5333*
提案手法	$p = 7, q = 0.5$	0.5514	0.5106	0.4151	0.7170	0.6538
提案手法	$p = 7, q = 0.6$	0.6476	0.7308	0.3750	0.7407	0.6964
提案手法	$p = 7, q = 0.7$	**0.7161**	**0.7170**	**0.5714**	**0.7843**	**0.7931**
提案手法	$p = 7, q = 0.8$	0.5482	0.4898	0.4490	0.7442	0.6667
提案手法	$p = 7, q = 0.9$	0.3693	0.3243	0.2286	0.5000	0.4557
提案手法	$p = 7, q = 1$	*0.2900*	*0.1429*	*0.2500*	*0.4138*	*0.3824*

の場合には，p による値の変化はほとんど見られない．

さらに，表 4.5 に 10 のテストセットの平均および個別のテストセット (murder, terrorism, scandal, s+k) を用いた場合のトピック抽出の F 値を示す．示した個々のテストセットについてみると，一つのトピックを示す文書数[*6]が少ないデータ (murder, terrorism 等) の方が，一つのトピックを示す文書数が比較的多いデータ (scandal, s+k 等) と比較して，q を変化させた場合の精度の変分が大きい．この傾向は，p が小さい場合

[*6] 表 4.3 の「ラベル付けされた文書数」を「ラベル数」で除算した値．murder:3.77, terrorism:3.75, scandal:7.73, s+k:5.13

($p = 3$) に顕著に現れ，一定以上 p が大きくなる ($p = 7$) と，テストセットによらず，「不要なエッジによるコアノードの埋没」が起きるため，全体的に，エッジ除去の効果が高くなる．

また，q を過度に下げた場合に再現率が低下する原因は，エッジを除去し過ぎたことによって，グラフ構造が必要以上に疎になったためであると考えられる．

一方，表 4.5 に示した比較対象手法のうち良い結果を出した凝集法と比較すると，$q = 1$ の場合 (斜体で表記) の場合，ほとんどのテストセット，p の条件で比較対象手法より精度が低い．それに対して，不要なエッジの除去処理を入れる事により提案手法の精度は向上し，$q = 0.7$ の場合 (太字で表記) には，p の値にかかわらずほとんど全ての条件で比較対象手法を上回ることが確認できた．以上の結果と，凝集法の結果が理想的なクラスタサイズを与えられた場合であるという事を考え合わせると，提案手法におけるトピック抽出精度は，比較対象手法を利用した場合より有益であると言える．

新聞記事を利用したクラスタリングの評価

4.2.3 で示したように，前節で特定した個々のコア文書に関係する文書 (サプリメンタル文書，サブトピック文書) を集め，1 つのクラスタに見立てることで，クラスタリングと同等の効果を得ることができる．特に本手法の一つの特徴として，文書集合中で関連性が高い部分に注目し，関連性が疎な部分を無視する点があり，今回用意したコーパスのように，その他 (1 文書でしか触れられないトピック) を多く含む文書集合のクラスタリングが高い精度で行えるのではないかと考えられる．

提案手法の実験条件は，トピック抽出と同じく表 4.4 に示した条件の全ての組合せで行った．

比較対象も，トピック抽出と同じく，K-means 法および凝集法 (重心法) を利用し，実験に利用するクラスタ数もトピック抽出と同じ理想的なクラスタ数とした．

評価指標としては F-Score[73] を利用した．F-Score はクラスタリングの精度を測定する 1 つの指標であり，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．算出式を以下に示す．

$$F - Score = \sum_{c \in C} \frac{|m_c|}{|m|} \max_{r \in R} \frac{2 \times |m_{r,c}|}{|m_r| + |m_c|} \tag{4.7}$$

ここで，C は正解データ中のクラスタの集合，c は正解データ中の個々のクラスタを示し，m_c はクラスタ c 中の文書の集合を示す．また，R はクラスタリング結果中のクラスタの

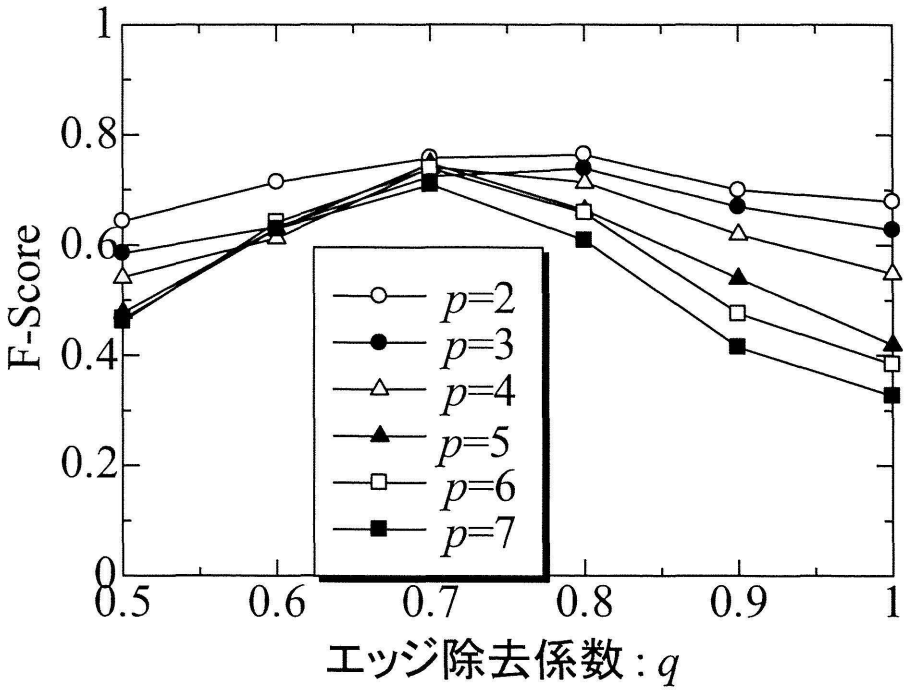

図 4.6 不要エッジ除去係数 (q) とクラスタリング精度の関係

集合，r はクラスタ結果中の個々のクラスタを示し，m_r はクラスタ r 中の文書の集合，$m_{r,c}$ は，m_r と m_c の積集合，m は正解データ中に含まれる文書の集合を示す．

また，サブトピックノードの収集に関する予備実験により，コアノードおよびサプリメンタルノードへの遷移確率が 50%(この値は 4.2.3 のサブトピックノードの定義に由来する) を多少下回るノードでもそのトピックに関連する場合が多いことが確認できたため，サプリメンタルノードおよびサブトピックノードへの遷移確率が 30% を上回るノードをサブトピックノードとした．

評価結果を図 4.6 に示す．グラフ上の全ての点は 10 個のテストセットの結果の平均である．全体的な傾向は，前節で述べたトピック抽出の傾向と類似しており，p の値にかかわらず $q = 0.7$ 前後で精度のピークを迎えており，その場合の F-Score はほぼ同等である．トピック抽出の F 値と F-Score の相関係数は 0.8829 であり，高い相関関係があると言える．この傾向は，コアノードを中心に関連する文書をたどる事で，クラスタを特定する提案手法の特徴を示している．

図 4.7 に，$q = 1$(エッジの除去を行わない条件) と $q = 0.7$(エッジ除去を行う条件) で，1 ノード辺りの外向きエッジ数 p を変化させた場合の F-Score の推移を示す．$q = 1$ の場合，p の増加により精度が低下しているが，$q = 0.7$ の場合には，ほとんど精度は変化

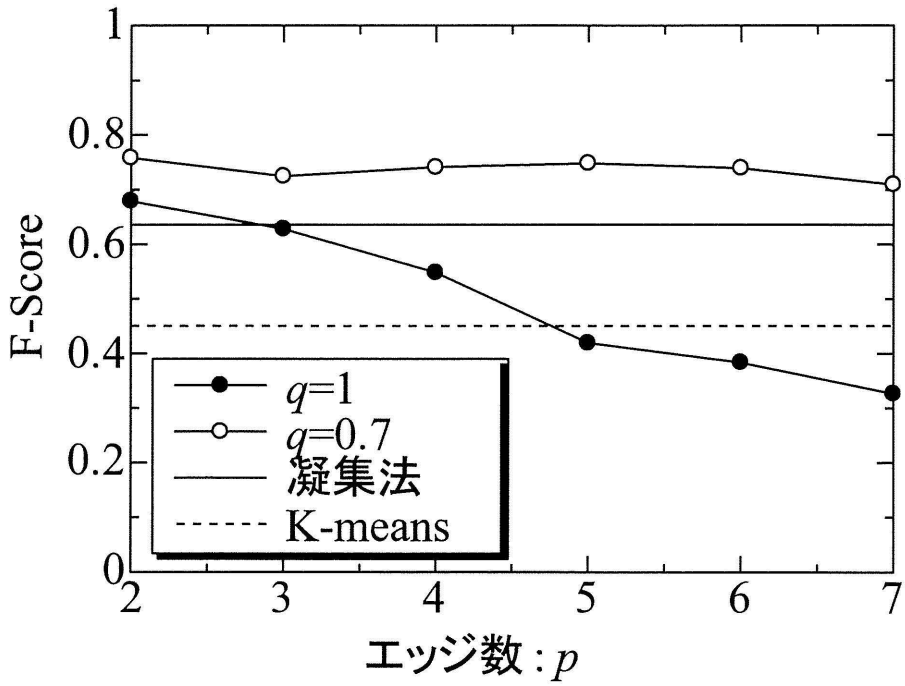

図 4.7 エッジ数 (p) とクラスタリング精度の関係 ($q = 1$, $q = 0.7$ の場合)

しないことがわかる．これは前述の通り，トピック抽出の精度と同じ傾向を示しており，エッジ除去を適切に行うことで，外向きエッジ数のパラメータ p には関係なく，精度が安定している事を示している．その他の傾向もトピック抽出と同傾向にあり，図 4.6 に示すとおり，p の値が小さい方が q の変化による精度の変化が小さいが，精度のピークは p の値によらない．また，表 4.6 に示すように，個々のデータを見た場合にも，p が小さい場合 ($p = 3$) には，一つのトピックを示す文書数が少ないデータ (murder, terrorism 等) の方が，一つのトピックを示す文書数が比較的多いデータ (scandal, s+k 等) と比較して，q を変化させた場合の精度の変分が大きい傾向にある．

また，図 4.7 には，比較対象とした凝集法と，K-means 法を用いた場合の精度も示している．比較的精度のよかった凝集法と比較しても，適切なエッジ除去を行った条件 ($q = 0.7$) では，その精度を上回っている事がわかる．ただ，凝集法では，類似した文書同士をクラスタとして統合していくため，理想的なクラスタ数が与えられた場合には，少ない文書から構成されるトピックのクラスタを高精度に取得する傾向があり，比較的高い F-Score を記録した．一方，K-means 法は今回のデータではあまり高い精度を示さなかった．

表 4.6 クラスタリング精度の評価結果 (F-Score)

手法	実験条件	average	murder	terrorism	scandal	s+k
提案手法	$p = 3, q = 0.5$	0.5852	0.5795	0.3512	0.7381	0.6808
提案手法	$p = 3, q = 0.6$	0.6329	0.6337	0.3903	0.7592	0.7218
提案手法	$p = 3, q = 0.7$	0.7243	0.7232	0.5988	0.7818	0.7840
提案手法	$p = 3, q = 0.8$	0.7384	0.8081	0.5677	0.7984	0.7903
提案手法	$p = 3, q = 0.9$	0.6695	0.6656	0.5731	0.7855	0.7333
提案手法	$p = 3, q = 1$	0.6278	0.6127	0.4903	0.7855	0.7200
提案手法	$p = 7, q = 0.5$	0.4659	0.3999	0.2998	0.6721	0.5631
提案手法	$p = 7, q = 0.6$	0.6293	0.6210	0.4376	0.8130	0.7213
提案手法	$p = 7, q = 0.7$	0.7092	0.6786	0.6321	0.7869	0.7817
提案手法	$p = 7, q = 0.8$	0.6082	0.5681	0.4770	0.7880	0.7229
提案手法	$p = 7, q = 0.9$	0.4148	0.3692	0.3325	0.6131	0.5031
提案手法	$p = 7, q = 1$	0.3268	0.1753	0.2900	0.5497	0.3988

NTCIR-4 WEB D に基づくクラスタリングの評価

提案手法によるクラスタリングを，別の指標において評価するため，NTCIR-4 で行われた「トピック分類タスク」[9] に基づく評価を行った．

このタスクの評価では，キーワード検索の結果をクラスタリングし，クラスタリングの結果得られたクラスタによってどの程度効果的に適合文書を探せたかという指標によって評価を行っている．具体的には，まず得られたクラスタを適合ページを含む量に応じて降順にソートし，最上位にランクされたクラスタから文書を n 件 ($n = 20$) 取得する．一つのクラスタに n 件文書が存在しない場合には次のクラスタを探索するというものである．

評価には rigid と relax の 2 種類があるが，これはどのような検索結果を適合文書とみなすかによる違いである．元々の適合性判定データには，各検索課題毎に，各文書に対して S(高適合)，A(適合)，B(部分的に適合)，C(不適合) の判定がつけられている．rigid ではこのうち S，A のみを，relax では S，A，B を適合文書とみなして評価を行う．

この評価では，それぞれの Web ページ全体を利用して作成したキーワードベクトルを用いて提案手法を適用した．ここでは，比較評価を行うことが目的であるため，$p = 3, 5$ と $q = 0.5$〜1 の組合せで実験を行った．また，提案手法ではすべての検索結果をクラス

表 4.7 NTCIR-4 WEB D の手法による評価結果 (rigid 条件の場合)

手法	実験条件	MAP	P@20	R@20
提案手法	$p = 3, q = 1$	0.2341	0.3864	0.3689
提案手法	$p = 3, q = 0.9$	0.3786	0.4182	0.4321
提案手法	$p = 3, q = 0.8$	**0.4358**	**0.4500**	0.5495
提案手法	$p = 3, q = 0.7$	0.2843	0.3818	0.3532
提案手法	$p = 3, q = 0.6$	0.3131	0.3545	0.3195
提案手法	$p = 3, q = 0.5$	0.2071	0.2318	0.2071
提案手法	$p = 5, q = 1$	0.2038	0.3500	0.4579
提案手法	$p = 5, q = 0.9$	0.3218	0.4455	0.6026
提案手法	$p = 5, q = 0.8$	0.3753	0.4227	0.4592
提案手法	$p = 5, q = 0.7$	0.2835	0.3818	0.3623
提案手法	$p = 5, q = 0.6$	0.1646	0.2273	0.1647
提案手法	$p = 5, q = 0.5$	0.1250	0.1682	0.1252
METAL	–	0.36	0.449	**0.754**

タ化するわけではなく，クラスタと見なさない文書 (アウトライヤー文書) も出力する．評価においては，このアウトライヤー文書は除去し，クラスタ化された文書だけを用いて評価を行っている．この前処理により，クラスタ化されない正解が存在した場合にクラスタリング精度とは関係なく評価値が上昇する事を防ぐ．

評価結果を表 4.7 および表 4.8 に示す．q の変化による結果の違いに着目すると，q を 1 から小さくしていくと，$q = 0.8$ 辺りで精度のピークを記録し，その後，低下するという傾向が見られた．ピーク時の結果を見ると，特に平均適合率 (MAP) の上昇は大きく，q を適正にセットすることで，上位ランキングのクラスタに多くの適合文書を含む．

また，NTCIR-4 において最も高い精度を示した METAL[47] との比較を行うと，regid および relax の両方の条件とも平均適合率および検索結果上位 20 件での適合率 (P@20) は提案手法が高い値を記録した．一方，検索結果上位 20 件での再現率 (R@20) は METAL の方が高い．この結果から提案手法はランキング上位に正解文書を集め易いが，必ずしも網羅的に正解を収集しないと言える．

この再現率が低い理由としては，提案手法ではクラスタ化できずにアウトライヤーノードと見なされた正解文書が存在し，前処理によって除去されている部分が大きい．この現

表 4.8 NTCIR-4 WEB D の手法による評価結果 (relax 条件の場合)

手法	実験条件	MAP	P@20	R@20
提案手法	$p = 3, q = 1$	0.2266	0.4545	0.4024
提案手法	$p = 3, q = 0.9$	0.2870	0.4818	0.3683
提案手法	$p = 3, q = 0.8$	**0.396**	**0.5364**	0.4730
提案手法	$p = 3, q = 0.7$	0.2535	0.4591	0.2794
提案手法	$p = 3, q = 0.6$	0.2324	0.4091	0.2396
提案手法	$p = 3, q = 0.5$	0.1297	0.2545	0.1349
提案手法	$p = 5, q = 1$	0.1681	0.3773	0.3335
提案手法	$p = 5, q = 0.9$	0.2860	0.4636	0.4889
提案手法	$p = 5, q = 0.8$	0.3498	0.5000	0.4069
提案手法	$p = 5, q = 0.7$	0.2172	0.4318	0.2648
提案手法	$p = 5, q = 0.6$	0.0995	0.2591	0.1075
提案手法	$p = 5, q = 0.5$	0.0528	0.1773	0.0564
METAL	–	0.3	0.48	**0.532**

象が起こる原因としては，一部の Web ページを処理する際に文書間の類似度が適切に評価できなかったのではないかと考えられる．つまり，今回の実験では Web ページ全体を使ってキーワードベクトルを作成したが，実際のデータ中には複数のトピックを含む文書も存在しており，そのような場合に類似度があいまいになった可能性がある．

基本特性評価のまとめ

以上では，2 つのコーパスを利用して，提案手法をトピック抽出，文書クラスタリングの観点から評価し，提案手法について分析した．

評価全体を通して提案手法はトピック抽出，クラスタリングの両方の観点から，比較対象と比べて良好な結果を得た．提案手法はトピック構造のマイニングを行う手法であるため，この評価が全てではないが，今回評価を行った指標はトピック構造のマイニングを行うために重要な要素であり，この評価結果により，提案手法の基本的な特性の有効性が証明できた．

新聞記事を利用した評価では，単純に決められた本数の外向きエッジを張った場合 ($q = 1$ の場合) には，不要なエッジがコア文書の抽出を妨げており，これが全ての精度を

下げている事が確認できた．この対策として，関連性の薄いエッジを除去することによって，この問題が解決する事がわかった．また，$q = 0.7$ 程度とすることで，p の値にかかわらず，トピック抽出，クラスタリングとも高い精度を示した．さらに，トピック抽出と，文書クラスタリングの結果の比較から，トピック抽出の再現率が文書クラスタリングの精度に影響する事がわかった．

また，NTCIR-4 の評価では，最新の手法との比較を行い，3 つの内 2 つの評価基準で，提案手法が最新の手法と比較して良好な結果を得た．ただし，今回の実験では，文書間の類似度を算出する場合に，文書全体の情報を利用しており，これによって多少精度が低下している可能性がある．この部分を，検索クエリー近傍の情報を利用して文書間の関係の評価を行う等の前処理を行うことで精度の改善が期待できると考えている．

4.3.3 トピック構造に関する仮説の検証

本節では，4.2.3 で定義した文書のタイプ別けが仮説通りに機能しているかについて検証する．本節の実験では前節で利用した 10 のテストセットを用い，前節の実験において，高いクラスタリング精度を示した条件 $(p = 5, q = 0.7)$ で処理を行った結果を利用した．

コア文書とトピックの主な内容の網羅性

ここでは，4.2.3 で定義したコア文書が，クラスタ中のトピックの中心的な内容を表現する文書であるかどうかについて検証した結果を示す．評価では，以下に定義する「主要トピック内容網羅率」を用いる．

「主要トピック内容網羅率」とは，各クラスタ中で多くの文書 (今回は過半数の文書) に存在する単語を主要トピック内容構成語と考え，選択された各文書もしくは複数の文書で，その文書が所属するクラスタの主要トピック内容構成語をどの程度網羅するかを示す値である．この値によって文書もしくは複数の文書が，クラスタ内の主要な内容をどの程度カバーしているかを示す．ただし，主要トピック内容構成語を抽出する際に，コーパス全体での出現が極端に高い語は，あらかじめ除去するものとする[*7]．

評価においては，クラスタ内に一定以上の文書が存在する場合でなければ，文書の選択に文書の種別を利用する意味はないと思われる事から，上記の処理結果からクラスタサイズが 5 以上のクラスタに含まれる文書のみを選択し，評価に利用した．

[*7] 今回の実験では，実験に利用したコレクション (毎日新聞 94, 95 年; 約 20 万件) で，文書頻度 (DF) が 5,000 以上の語をあらかじめ除去した．

表 4.9 コア文書の主要トピック内容網羅率

文書タイプ	主要トピック網羅率
コア文書	0.9072
全ての文書	0.7703
コア文書以外の文書	0.7520

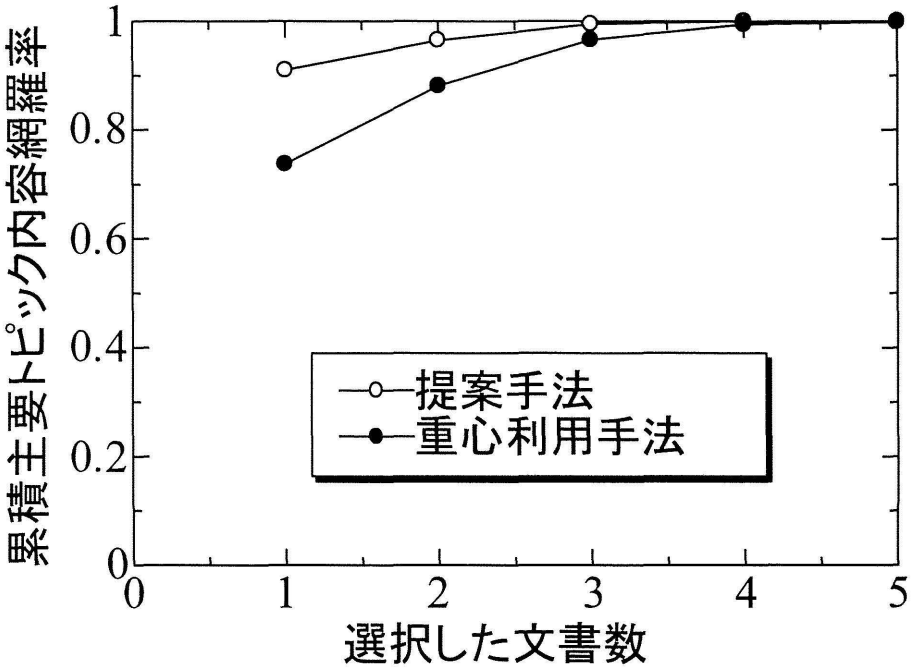

図 4.8 主要トピック内容網羅率の推移

表 4.9 に，すべてのコア文書の主要トピック内容網羅率の平均と，全ての文書の主要トピック内容網羅率の平均，コア文書を除く文書の主要トピック内容網羅率の平均を示す．

この結果から，コア文書を選択した場合には，主要トピック内容の 90% 以上を網羅している事，それ以外の文書では平均して 75% 程度と差が存在することがわかる．

次に，提案手法以外の手段によって文書を選択した場合との比較実験を行った．この実験では，提案手法および比較手法によって，文書をいくつか選択した場合の各時点で，主要トピック内容網羅率がどのように遷移するかを検証する．

提案手法を利用した場合には，文書をコア文書，サプリメンタル文書，サブトピック文書の順に選択する．同じタイプの文書は，より中心性スコアが高い文書を先に選択する事とした．この順に文書を選択することで，早い段階で主要トピック内容網羅率が 100% に

近付く事を想定している.

比較対象として，クラスタの重心に近い文書から順に選択する手法を利用した．この手法では，対数化 TF-IDF 重みを持ったキーワードベクトルで各文書を表現し，距離の算出には，コサイン類似度を利用した.

実験では，前記の実験と同様に，提案手法で作成されたクラスタのうちクラスタサイズが 5 以上のものを選択し，二つの手法による文書選択の差を評価した.

結果を図 4.8 に示す．横軸は選択した文書数を示し，縦軸は累積主要トピック内容網羅率を示す．つまり，X 軸の値が 2 の時のプロットは，文書を 2 つ選択した場合，2 つの文書で主要トピック内容をどの程度網羅しているかを示している．全てのプロットはすべてのクラスタについて算出した値を平均している.

結果を見ると，1 文書目，2 文書目を選択した段階で，明らかに提案手法の方が高い網羅率を示している．これは，提案手法を利用した場合が，比較手法を利用した場合と比べて，より少ない文書で，主要なトピックを取得する事ができる事を示している．累積値であるため，その後の差は小さくなるが，文書選択の初期の段階で差が明らかとなった.

重心を利用した比較対象手法の網羅性が高くならなかった原因としては，IDF の高い語の影響が考えられる．IDF が極端に高い語を持つ文書は，重心と近い文書と判断され易いが，そのような文書は今回のように主要なトピックの内容を選択するには不適切であったと考えられる．同様の傾向は，重要文抽出のタスクにおいても報告されている [10].

以上によって，提案手法によるコア文書が，他の文書と比較して，クラスタ中の主要な内容を選択すること，クラスタの重心との距離を利用した手法と比較して提案手法による優先付けが，クラスタ内の主要なトピックを収集するのに有益である事が検証できた.

サプリメンタル文書について

ここでは，4.2.3 で定義したサプリメンタル文書について検証を行う．4.2.3 では，サプリメンタル文書の定義として，コア文書と双方向エッジで直接もしくは間接的に接続されたノードであるとした．しかし，コアノードと直接接続しているノード (直接サプリメンタルノードと呼ぶ) と，間接的にしか接続していないノード (間接サプリメンタルノードと呼ぶ) にはその性質に差があるのではないかと考えられるため，その差位について検証した.

本節での検証では，前節で利用した「主要トピック内容網羅率」に加えて，以下で定義する「希少トピック内容網羅率」を利用している.

ここで，「希少トピック内容網羅率」とは， 各クラスタ中で少ない文書 (今回は 2 文書

表 4.10 各文書タイプにおける平均主要トピック内容網羅率，平均希少トピック内容の網羅率

文書タイプ	主要トピック内容網羅率	希少トピック内容網羅率
コア文書	0.9575	0.0698
直接サプリメンタル文書	0.8460	0.0990
間接サプリメンタル文書	0.7290	0.1176
サブトピック文書	0.6756	0.1559

とする) にのみ存在する単語を希少トピック内容構成語と考え，選択された各文書もしくは複数の文書で，その文書が所属するクラスタの希少トピック内容構成語をどの程度網羅するかを示す値である．この値によって文書もしくは複数の文書が，クラスタ内の希少なトピックの内容をどの程度どカバーしているかを示す．ただし，主要トピック内容構成語を抽出する場合と同様に，希少トピック内容構成語を抽出する際に，コーパス全体での出現が極端に高い語は，あらかじめ除去した[*8]．

実験では，コア文書，直接サプリメンタル文書，間接サプリメンタル文書，サブトピック文書のそれぞれについて，「主要トピック内容網羅率」，「希少トピック内容網羅率」を算出し，その差に有意性があるかどうかを t 検定で検証した．この評価には，前節の評価で用いたクラスタ (クラスタサイズが 5 以上) の中で，上記に示す 4 種のノードを全て含むクラスタを抽出し，評価に用いた．

表 4.10 に，結果を示す．「主要トピック内容網羅率」に関しては，コア文書と直接サプリメンタル文書の間 (両側検定: $t(46) = 9.70, p < 0.01$)，および直接サプリメンタル文書と間接サプリメンタル文書の間 (両側検定: $t(46) = 5.39, p < 0.01$)，間接サプリメンタル文書とサブトピック文書の間 (両側検定:$t(46) = 2.17, 0.01 < p < 0.05$) の全てのノード間に有意差があるとの結果が得られた．また，「希少トピック内容網羅率」に関しては，コア文書と直接サプリメンタル文書の間 (両側検定:$t(46) = 6.09, p < 0.01$)，および間接サプリメンタル文書とサブトピック文書の間 (両側検定:$t(46) = 3.31, p < 0.01$) に有意な差があるとの結果が得られた．

つまり，直接サプリメンタルノードと，間接サプリメンタルノードについては，「希少トピック内容網羅率」に関しては差がないものの，「主要トピック内容網羅率」に差がある．一方，間接サプリメンタル文書とサブトピック文書の間には両方の指標で差がある．

[*8] 今回の実験では，実験に利用したコレクション (毎日新聞 94, 95 年; 約 20 万件) で，DF が 5,000 以上の語をあらかじめ除去した．

表 4.11 サブトピック文書の希少トピック内容網羅率

文書タイプ	希少内容トピック網羅率
サブトピック文書	0.1787
全文書	0.1370
サブトピック文書以外の文書	0.1208

以上から，直接サプリメンタル文書は，より中心に近いトピックを含む文書であり，間接サプリメンタル文書は，中心的なトピックを扱う傾向はやや弱く，希少なトピックを扱う傾向が高くなるという結果となった．しかし，サブトピック文書との差は明確であることから，間接サプリメンタル文書は，他の文書にも出現するトピック内容を含む文書であると考えられる．

また，今回の検証により，4.2.3 で提案した文書のタイプ間に明確な差がある事を示すことができた．

サブトピック文書と希少トピック内容の網羅性

ここでは，4.2.3 で定義したサブトピック文書が，他の文書と比較してクラスタ中のトピックのノベルティの高い情報 (クラスタ内で希少性が高い内容) を含む文書であるかどうかについて検証した．

この検証では，前節で利用した「希少トピック内容網羅性」を利用する．本節の実験では，クラスタサイズが 5 以上のクラスタのうち，コア文書，サプリメンタル文書，サブトピック文書の全てが存在するクラスタのみを選択し評価を行った．

表 4.11 に，すべてのサブトピック文書の網羅率の平均と，全ての文書の網羅率の平均，サブトピック文書を除く文書の網羅率の平均を示す．サブトピック文書を選択する事で，ランダムに選択する場合と比較して希少トピック内容網羅率が約 30%(0.1370 から 0.1787) 上昇することがわかる．

次に，提案手法以外の手段によって文書を選択した場合との比較実験について示す．この実験では，提案手法および比較手法によって，文書をいくつか選択した場合に，累積希少トピック内容網羅率がどのように遷移するかを検証する．

提案手法を利用した場合には，文書をサブトピック文書，サプリメンタル文書，コア文書の順に選択する．同じタイプの文書が存在する場合には，より中心性スコアが低い文書を先に選択する事とした．この順に文書を選択することで，早い段階で希少トピック網羅

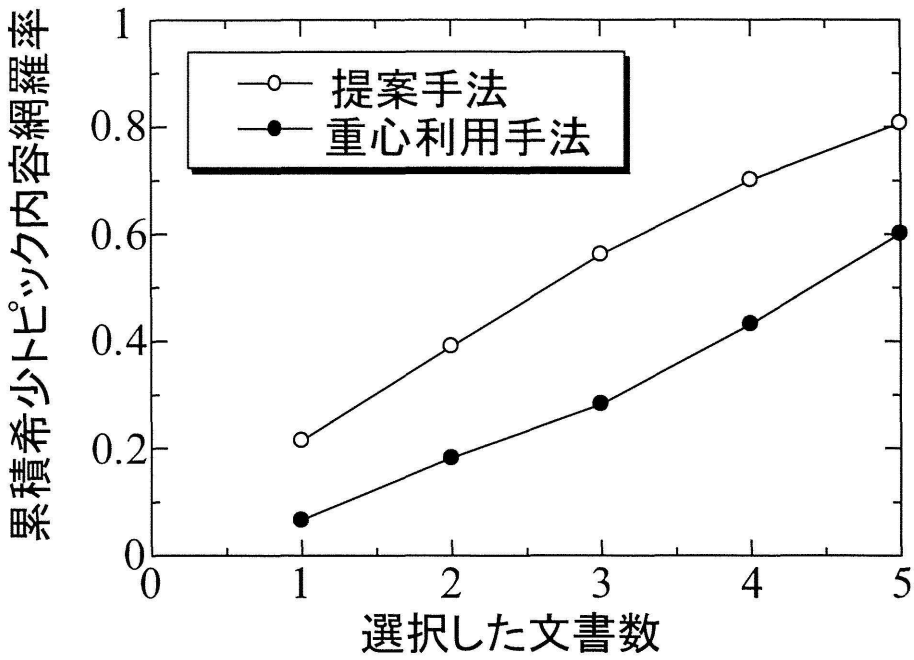

図 4.9 希少トピック内容網羅率の推移

率が高くなる事を想定している.

比較対象として，クラスタの重心に遠い文書から順に選択する手法を利用した．この手法では，対数化 TF-IDF 重みを持ったキーワードベクトルで各文書を表現し，距離の算出には，コサイン類似度を利用した.

実験では，提案手法で作成されたクラスタのうちクラスタサイズが 5 以上で，コア文書，サプリメンタル文書，サブトピック文書の全てが存在するクラスタのみを選択，二つの手法による文書選択の差を評価した.

結果を図 4.9 に示す．横軸は選択した文書数を示し，縦軸は累積希少トピック網羅率を示す．つまり，X 軸の値が 2 の時のプロットは，文書を 2 つ選択した場合，2 つの文書で希少トピック内容をどの程度網羅しているかを示している．全てのプロットはすべてのクラスタについて算出した値を平均している.

結果を見ると，全てのステップで明確な差がある事が確認できる．このグラフから同じ網羅性に到達するには，重心法を利用した場合と比較して，提案手法を用いると，文書の選択数が 1〜2 減少するということがわかる．つまり，提案手法で 1 文書を閲覧した場合と同程度の希少トピック内容を網羅する場合，重心法を利用した手法では，2 文書以上閲覧することが必要となる.

以上によって，提案手法によるサブトピック文書が，他の文書と比較して，クラスタ中

の希少なトピック内容をより多く含むこと，クラスタの重心との距離を利用した手法と比較して提案手法による優先付けが，クラスタ内の希少なトピック内容を収集するのに有益である事が検証できた．

4.4　可視化

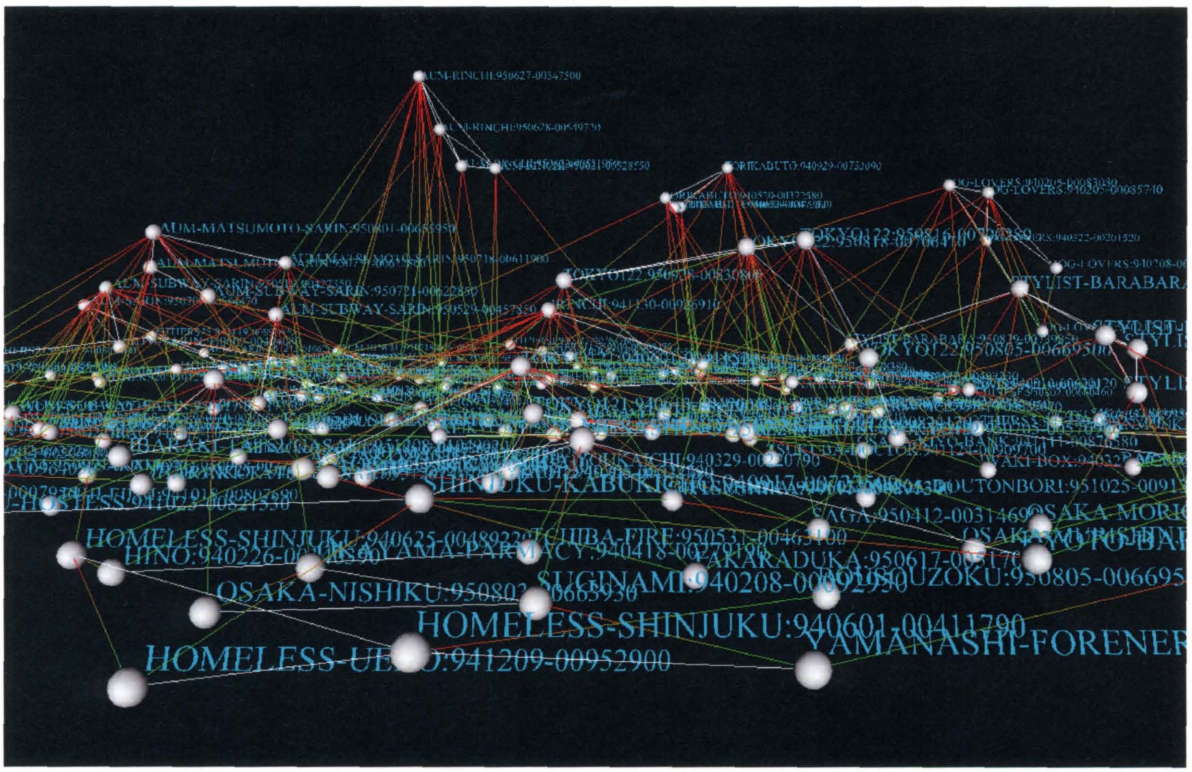

図 4.10　“murder” コーパスの可視化結果 ($p = 3, q = 1$)

ここではトピック構造を利用した文書集合のブラウジングおよびマイニングの有効性を示すため，文書集合グラフとそのノードの中心性スコアを利用した文書集合の可視化について示す．4.2.3 で述べた全てのノードを可視化している．

図 4.10 および図 4.11 にそれぞれ $p = 3, q = 1$ および $p = 3, q = 0.7$ で新聞記事の「murder」セットを可視化した結果を示す．この可視化において，グラフ構造は山田ら [61] の手法を利用して構成した．この手法では，エッジの有無を要素とする隣接行列と，2 次元平面上にプロットした距離を要素とする距離行列とのクロスエントロピーを最小化することによってノードを配置する．このグラフ構造を 3 次元空間の XY 平面上に配置し，Z 軸方向に中心性のスコアをプロットするという方法で，各文書およびその関係を 3

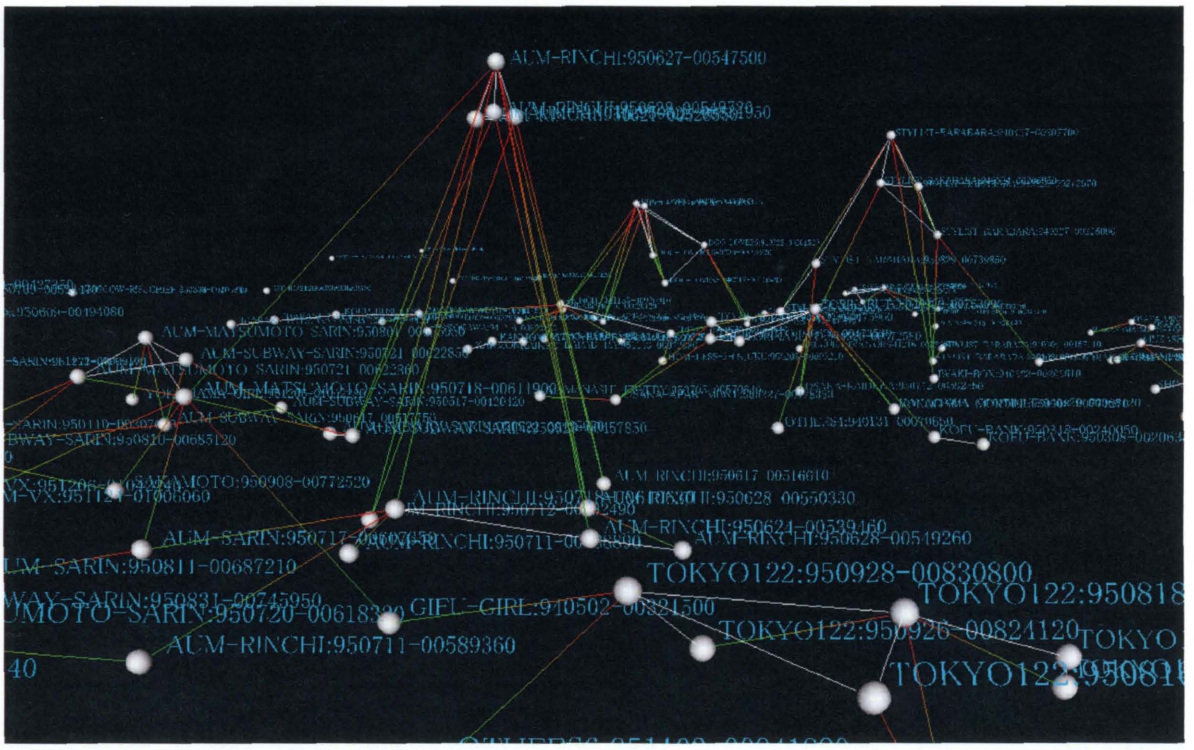

図 4.11 “murder” コーパスの可視化結果 ($p = 3, q = 0.7$)

次元空間上に表現している．白い球がノード (文書) を示し，球を繋ぐ線がエッジを示す．エッジの方向はグラデーションで表現されており，「from」側が黄緑色で，「to」側が赤色で表現されている．双方向エッジは白いラインで示されている．今回の例ではノードの横に淡い色で文書 ID と文書に付与されたラベルを表示しているが，ブラウジングやマイニング用のインタフェースとした場合には文書のタイトルを表示することが考えられる．

図 4.10,4.11 では，いくつかの山状の構造が見られる．それぞれの山はそれぞれ異なるトピックを表現しており，山の頂上に来る文書がコアノードである．これらの山状の構造から文書を選択することで，特定のトピックに関する所望の文書を発見したり，トピック間の関連を理解することができる．

図 4.11 に示す $q = 0.7$ の場合には，図 4.10 の下部に多く存在していたノードが減少している．これは，不要エッジの除去によって，エッジを失ったことで，他のノードとつながりがなくなり，単独で存在するように配置されたためである．二つを比べると，図 4.11 は明らかに疎な状態になっており，トピックを示す山も，関連性の薄い山同士は離れる等トピック間の関連性がより明確に可視化されている．

図 4.12 は図 4.11 の一部をズームして表示したものである．この図では 2 つのトピック

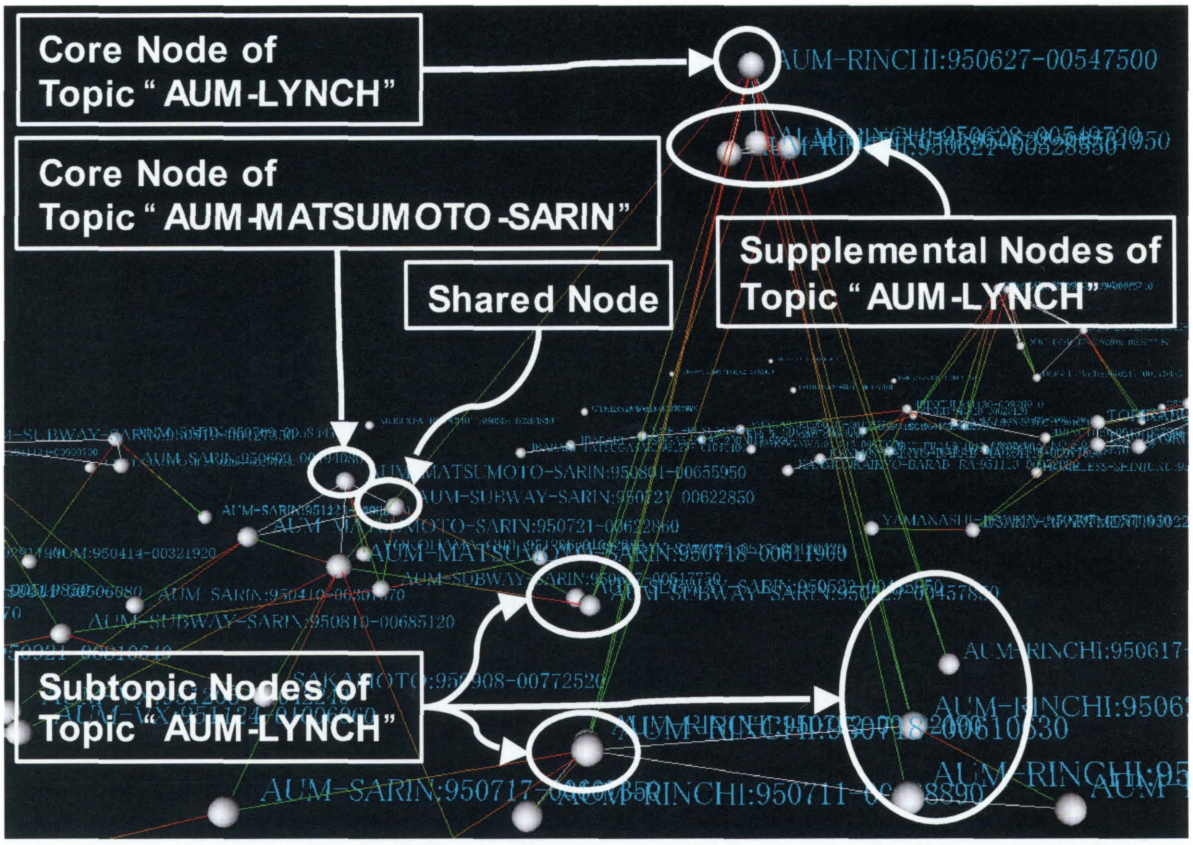

図 4.12　“murder” コーパスの可視化結果 (拡大版，$p = 3, q = 0.7$)

が高いレベルのノード (サプリメンタルノード) を共有している様子がわかる．これは 2 つのノードが強く関連していることを示している．実際にこれら 2 つのトピックは同じ宗教団体の起こした殺人事件について示している．また，この図では「AUM-LYNCH」のトピックに関するコアノード，サプリメンタルノード，サブトピックノードが見える．図 9 では，下部にアウトライヤーノードが多く存在いたため，不明瞭であったが，不要なエッジの除去により，主要なトピックに関するノードを見つけやすくなっている．

以上のトピック構造を利用した文書集合の可視化により，単に文書集合から存在するトピックやあるトピックに関係する文書群を特定するだけでなく，視覚的にトピック間のつながりや，個々の文書の位置付けをマイニングすることが可能であると言える．

4.5 タイムスタンプ付き文書に対するトピック構造マイニングの適用

以上で述べた手法では，文書間の類似度として文書の内容のみに着目している．しかし，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おり，このような文書を扱う場合には，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．

ここで，文書集合中に含まれるトピックを時間との関係の観点で分類すると，以下のように別けられる．

- 一過性のトピック
- 継続的なトピック
  - 周期的なトピック
  - ライフスパンが長いトピック

一過性のトピックとしては，ニュースで配信される事件や事故等が例として挙げられ，このようなトピックは特定のタイムスタンプに集中して関連する文書が生成される事が考えられる．また，周期的なトピックとしては，桜の開化等季節や暦に対応しておこるトピックが挙げられる．ライフスパンが長いトピックは，時期と関係なく継続的に起こるトピックであるため，タイムスタンプとの相関が小さい．

ここでは，上記に示すトピックのうち，タイムスタンプとの相関が大きいと考えられる一過性のトピックに注目した．これは，「湾岸戦争」の記事と「イラク戦争」の記事のように，内容的には類似しているが，時間的な近さを考慮すれば明確に見分けられる話題が多く存在すると考えたためである．

そこで，以下では，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，トピック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．

提案手法では，文書集合グラフを構築する段階で，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トピック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，4.5.1 で時間的近さを考慮する類似度およびそれを利用した文書集合グラフの構築について示し，4.5.2 で時間的近さを考慮することで，文書集合からのトピック抽出お

よびクラスタリングの精度がどのように変化するかについて示す.

4.5.1 時間的近さを考慮した文書集合グラフの構築

ここでは,時間的近さを考慮した文書集合グラフの構築を行うために利用する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,グラフ構造構築の際に利用することで,時間的近さを考慮したトピック構造マイニングを可能とする.

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.

$$\frac{d}{dt}TimeWeight(t) = -\lambda \times TimeWeight(t) \tag{4.8}$$

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.

$$TimeWeight(t) = T_0 \times \exp(-\lambda t) \tag{4.9}$$

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ため,以下のよう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.

$$TimeWeight(t) = T_0 \times \exp(-\frac{0.693}{t_{1/2}}t) \tag{4.10}$$

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 4.13 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.

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.

$$sim(i,j) = sim'(i,j) \times ((1-\alpha) + \alpha \times TimeWeight(t(i,j))) \tag{4.11}$$

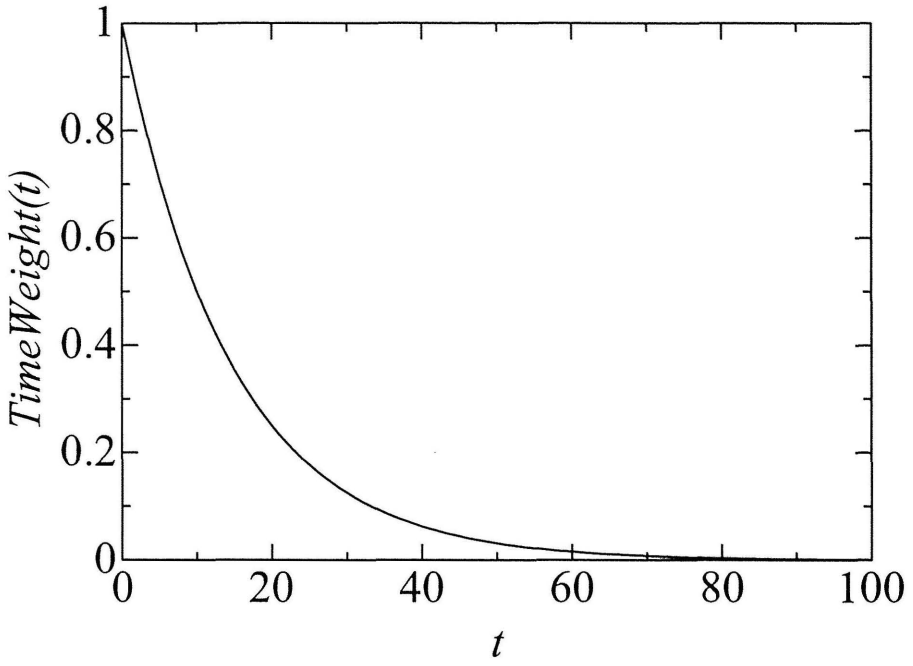

図 4.13 時間類似度の関数形状

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．$\alpha = 0$ の場合，類似度は通常の文書内容のみに基づく類似度となる．

図 4.14 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．左のグラフでは，内容に基づく類似度が同じ場合に，タイムスタンプの差の変化で，類似度 $sim(i,j)$ がどのように変化するかを示している．一方，右のグラフでは，タイムスタンプの差が同じ場合に，内容に基づく類似度が異なると，類似度 $sim(i,j)$ がどのように変化するかを示している．

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要素には，文書を形態素解析器 ChaSen[*9] 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[*10]．各要素の重みは対数化 TF-IDF で算出した．

従来のトピック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

[*9] http://chasen.naist.jp/hiki/ChaSen/
[*10] ただし，未知語の連続は一つの語として扱った．

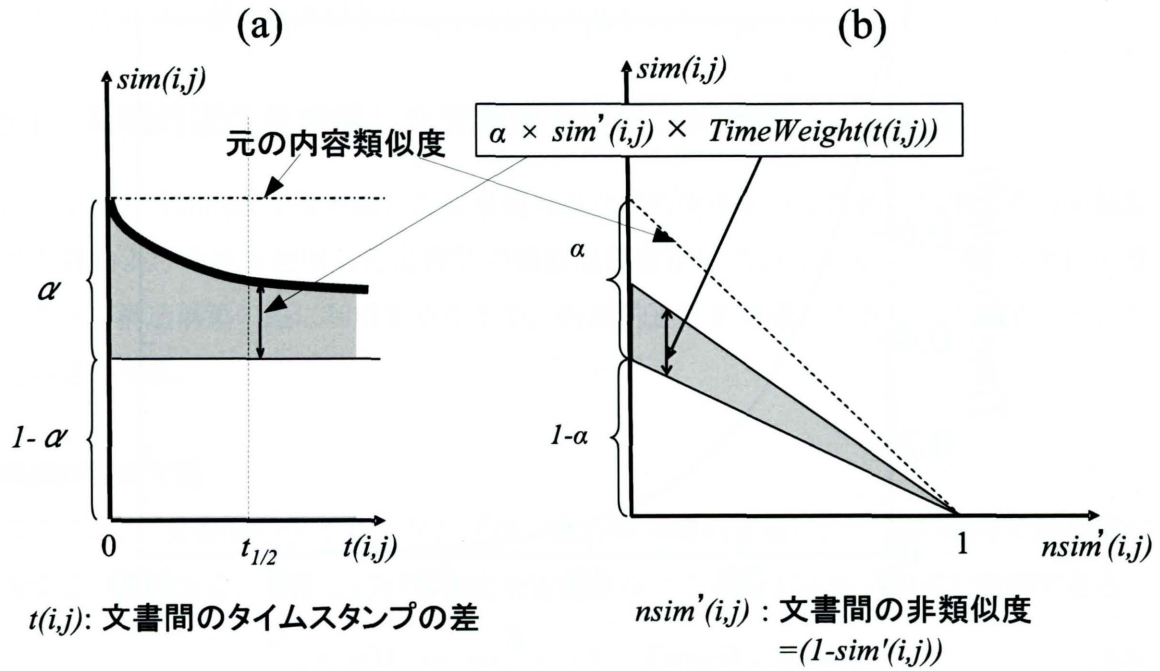

図 4.14 類似度の変化: (a) タイムスタンプの差が変化した場合, (b) 内容類似度が変化した場合

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トピック構造マイニングを行う．

4.5.2 基本特性の評価

本評価では，時間類似度を考慮することによるトピック抽出，クラスタリングの精度変化について明らかにし，その原因を考察する．ここでの評価では，4.3.2 で利用したコーパスおよび正解データを用いて，トピック抽出およびクラスタリングの評価を行う．

実験条件

ここで，パラメータ p と q は，4.2.1 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．p は各ノードからの外向きエッジの本数を規定する値であり，q は，関連性の薄いエッジを除去するためのパラメータである．今回の評価では，時間類似度を考慮しない場合に，トピック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比較対象とした．

また，時間類似度に関するパラメータとして，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．

以下に示すトピック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，4.3.2 と同等の手法で行った．

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．

トピック抽出精度について

まず比較対象と同じ条件 $(p = 5, q = 0.7)$ において，時間類似度を導入した場合の結果を図 4.15 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．本節で示すグラフ内の破線は比較対象である時間情報を考慮しない手法 $(p = 5, q = 0.7)$ の評価値を示す．

トピック抽出の適合率は，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，トピック抽出の精度が低下する事を示している．

しかし，今回の時間類似度の導入法の性格上，時間重みを強くした場合にも，内容が考慮されなくなるわけではなく，内容，時間の両面で類似した関係が強調されると考えられ，誤ったエッジが生成されたとは考えにくい．

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 4.16 に示す．比較対象手法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，エッジ数が減少している事がわかる．

このエッジ数の減少はトピック構造マイニングのエッジ除去プロセスが関係している．4.2.1 で示したように，エッジ除去プロセスでは，ノイズとなるエッジ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エッジを除去している．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，有益なエッジ同士でも重みの差が大きくなっている．このため，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，有益なエッジまでもが除去され，再現率が低下したのではないかと考えられる．

この有益なエッジの除去を抑える一つの方法は，不要エッジ除去係数 q の値を大きく設

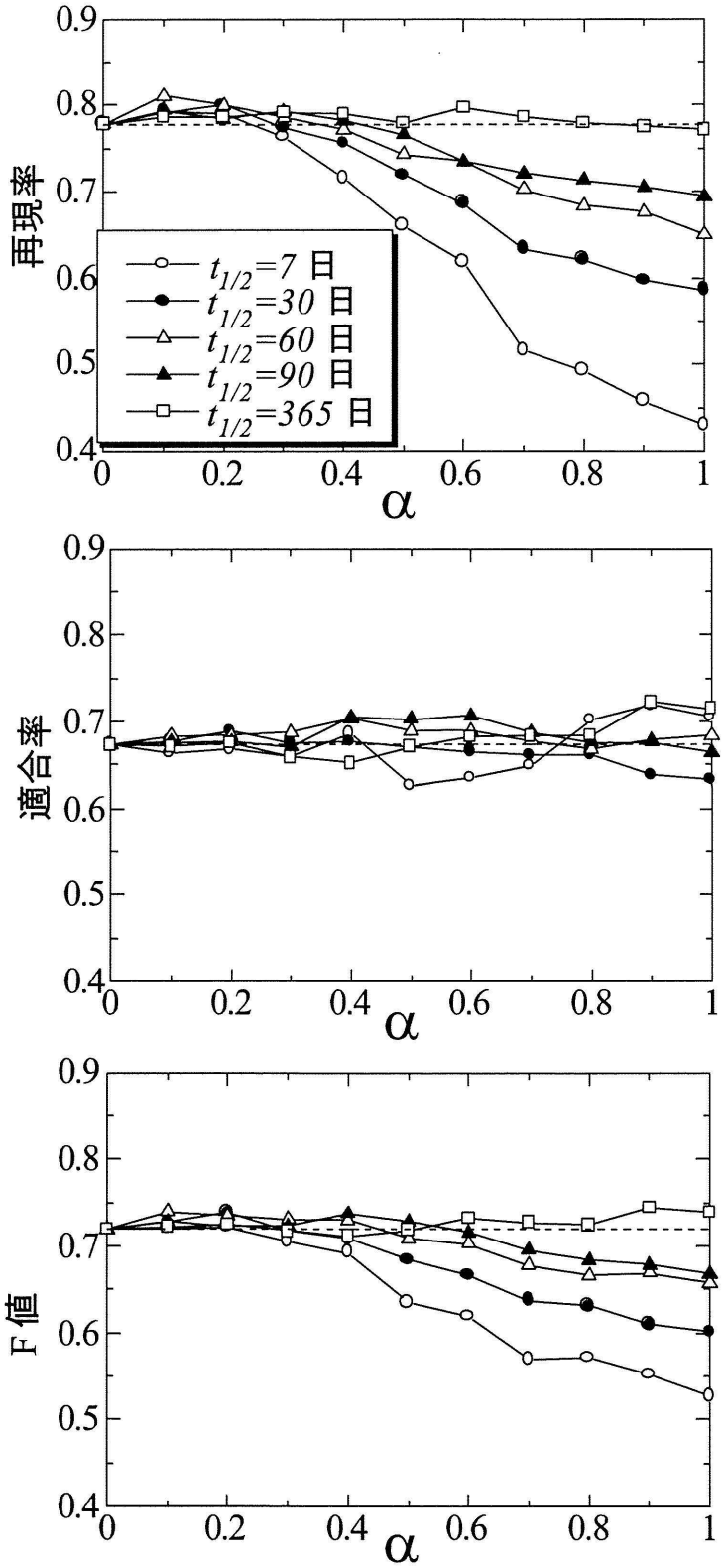

図 4.15 時間類似度に関するパラメータとトピック抽出精度の関係 ($p = 5, q = 0.7$)

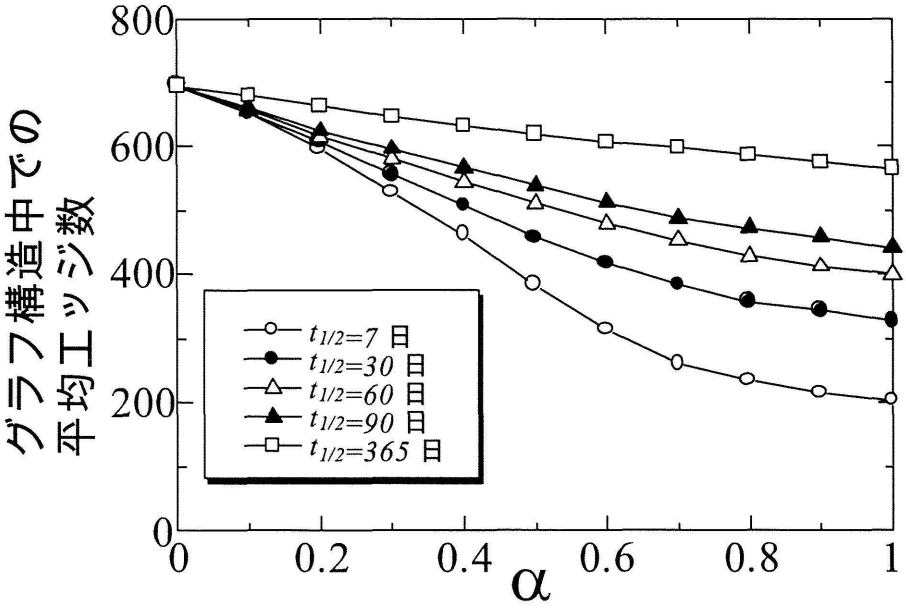

図 4.16 時間類似度に関するパラメータによるエッジ数の変化 ($p = 5, q = 0.7$)

定し，ある程度小さい重みのエッジも残す事である．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項 (式 (4.1) 中の $d_{max}E - D$) が高い割合を占めるようになり，自己遷移以外のエッジが除去される状況が考えられる．これを回避するには，1 ノード辺りのエッジ数 p を小さくし，式 (4.1) の d_{max} を小さくする事が考えられる．

以上を踏まえ，有益なエッジの除去を防ぐために，エッジの除去を少なくした条件 ($p = 5, q = 0.8$) および 1 ノード辺りのエッジ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 4.17-4.19 に示す．

“$p = 5, q = 0.8$” の場合 (図 4.17) 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，エッジ除去が十分に行われておらず，ノイズとなるエッジが残存し，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，グラフ構造中の総エッジ数が比較対象手法の 1.5 倍 (1054) である事，抽出される平均トピック数が比較対象手法より少ないこと (比較対象手法: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，不要なエッジが除去され，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，半減期が短い場合に比較対象手法の精度を上回った．ただし，F 値において，比較対象手法との間に有意な差はなかった．

“$p = 3, q = 0.7$” の場合 (図 4.18) は，再現率は元々比較対象手法を上回り，時間類似度

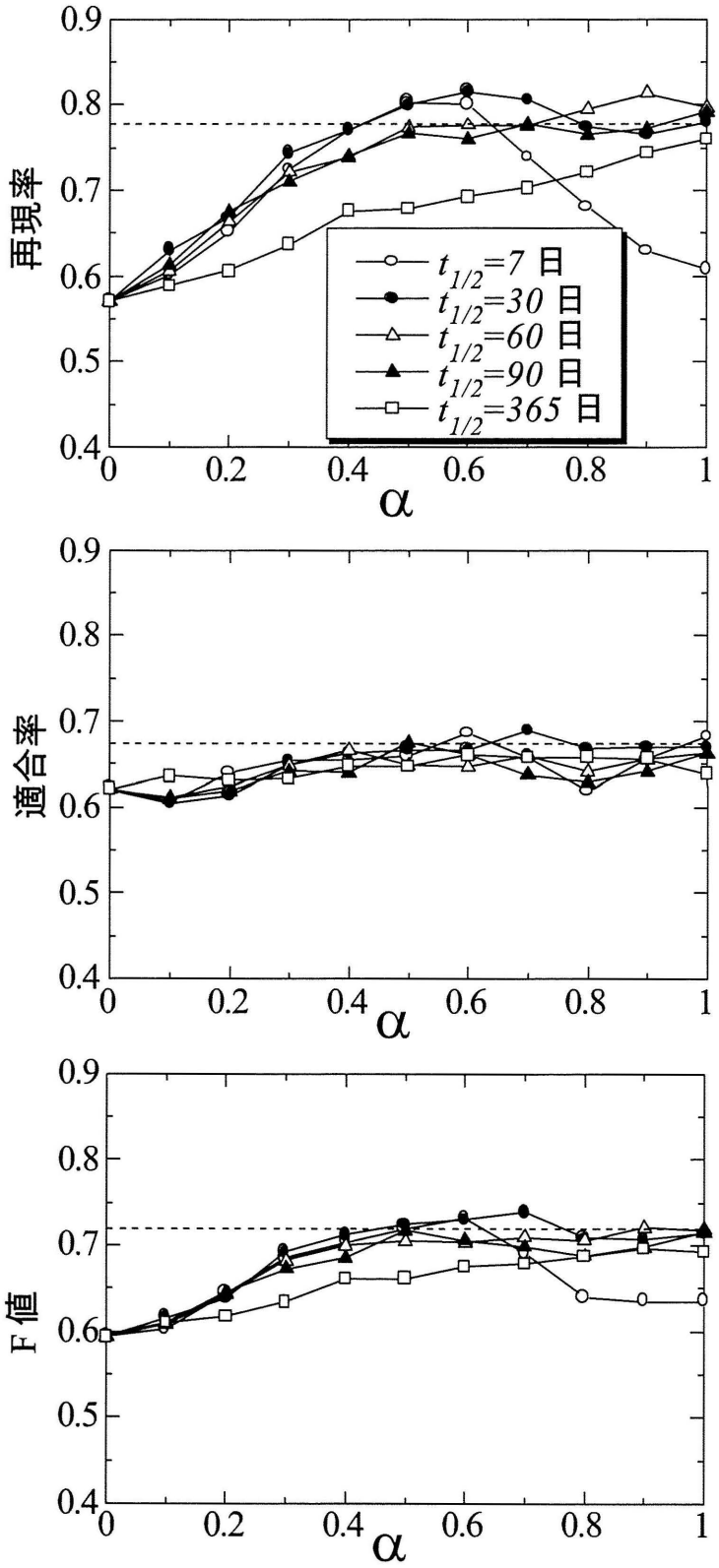

図 4.17 時間類似度に関するパラメータとトピック抽出精度の関係 ($p = 5, q = 0.8$)

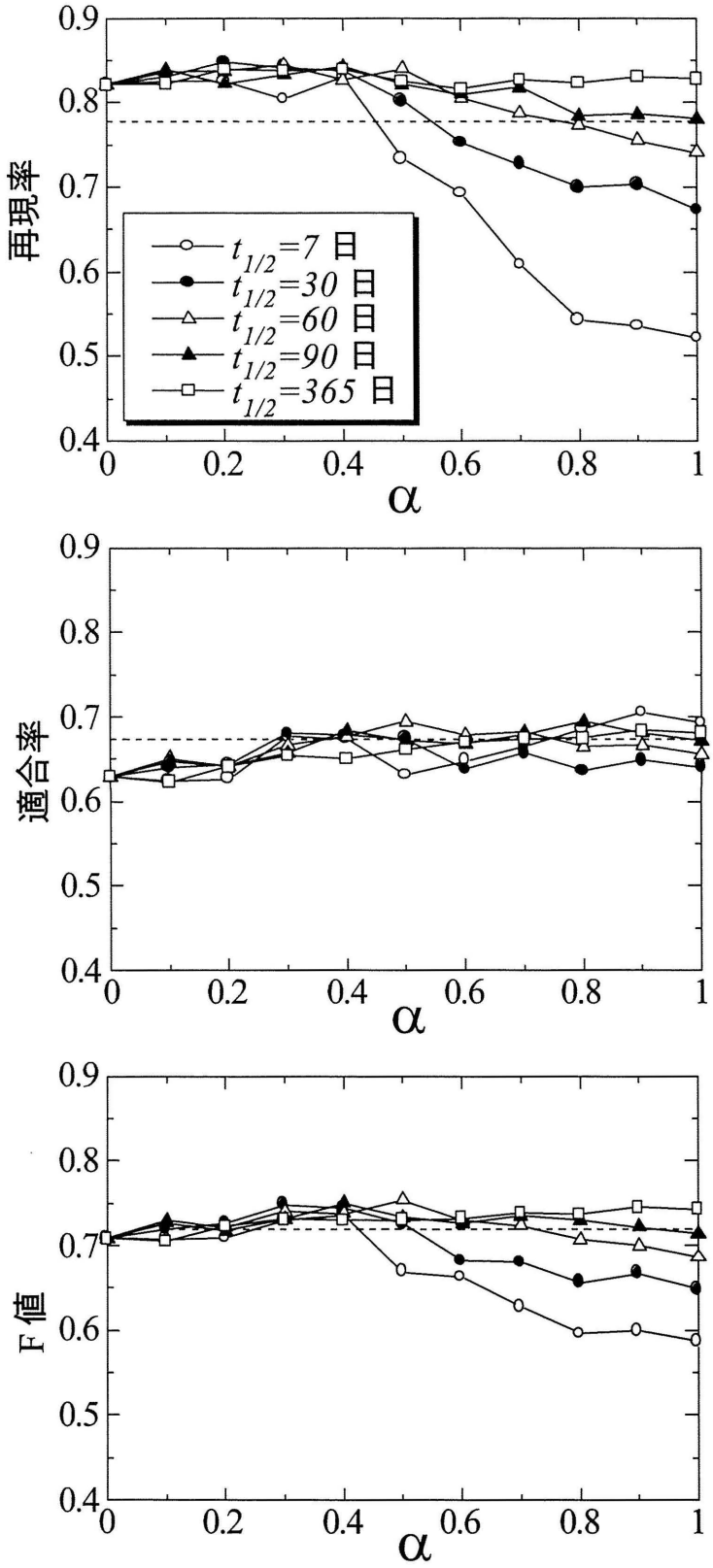

図 4.18 時間類似度に関するパラメータとトピック抽出精度の関係 ($p = 3, q = 0.7$)

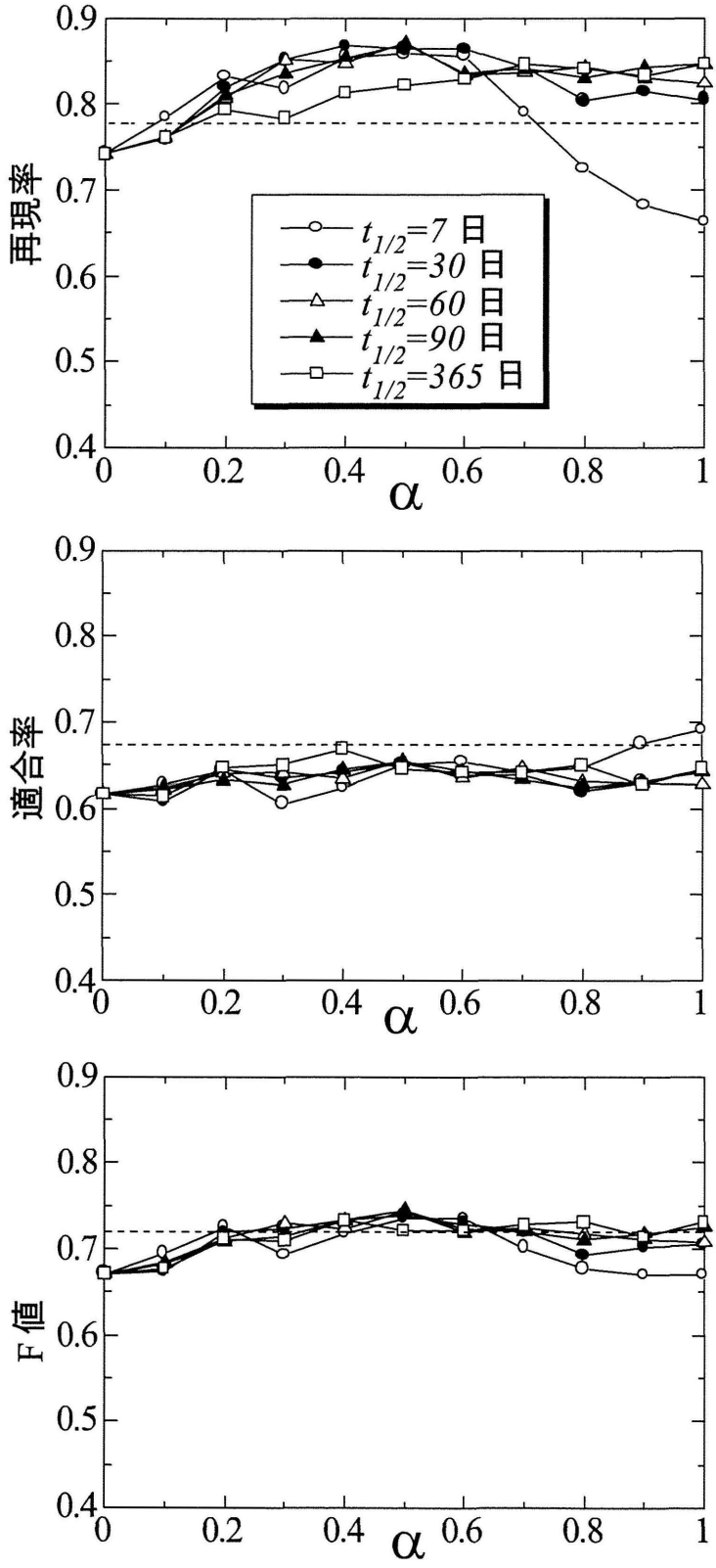

図 4.19 時間類似度に関するパラメータとトピック抽出精度の関係 ($p = 3, q = 0.8$)

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，比較対象手法を下回るが，時間類似度を導入する事で，比較対象手法に近い値を示す．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 値で比較対象手法を上回り，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エッジ数は "$p = 5, q = 0.7$" の場合と同程度少ないが，有益なエッジが残存したため，トピック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，抽出されるトピック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 (図 4.19)，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，比較対象手法と比較して最高 10% 程度向上している．これらの条件では，抽出トピック数が比較対象手法より 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，比較対象手法よりは低い．しかし，$t_{1/2} = 90, \alpha = 0.5$ の場合には，トピック抽出の F 値が 0.7443 程度を示し，t 検定の結果，比較対象手法の結果との間には有意な差があった (両側検定: $t(9) = 2.5723, .01 < p < .05$)．

クラスタリング精度

図 4.20 に比較対象手法の条件 ($p = 5, q = 0.7$) 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．トピック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．このようにトピック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トピック抽出で抽出するコアノードを元にしているためである．このため，図 4.16 で示したエッジ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，比較対象手法と比較して，20%〜50% 程度少ない (比較対象手法:190，正解データ:193.6)．

そこで，トピック抽出と同様に，エッジの除去を少なくした条件 ($p = 5, q = 0.8$) および 1 ノード辺りのエッジ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 4.21 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エッジ数が少ない条件である．

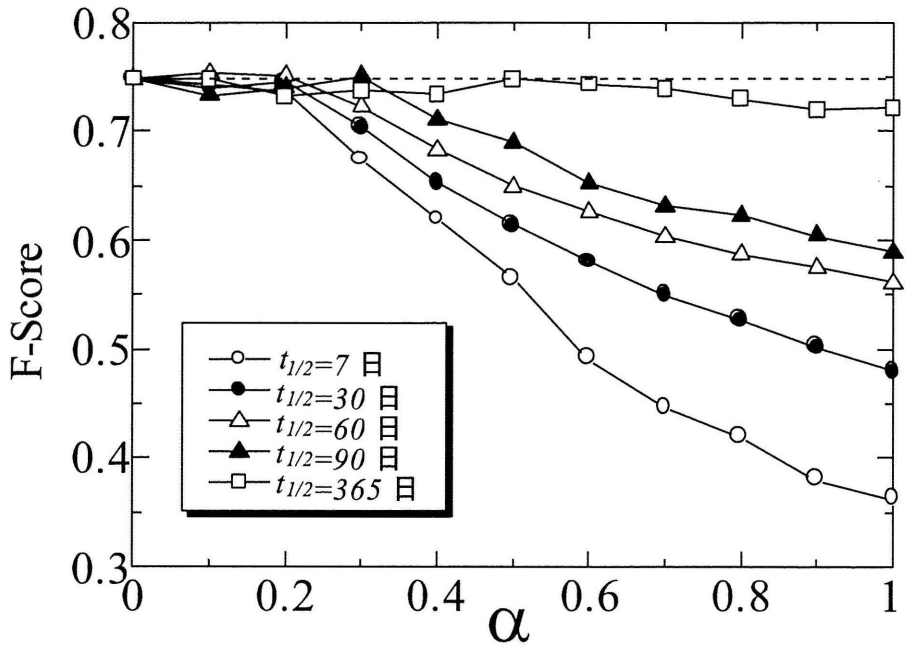

図 4.20 時間類似度に関するパラメータとクラスタリング精度の関係 ($p = 5, q = 0.7$)

エッジの質の違いで，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，比較対象手法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 $\alpha = 0.5$ 前後で比較対象手法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，比較対象手法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比較対象手法と同等以上であった．

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，トピック抽出およびクラスタリングの精度が比較対象手法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，外向きエッジ数 (p) を少なくし，かつ不要エッジ除去係数 (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，不要エッジ除去プロセスで有益なエッジが除去される傾向があるためである．

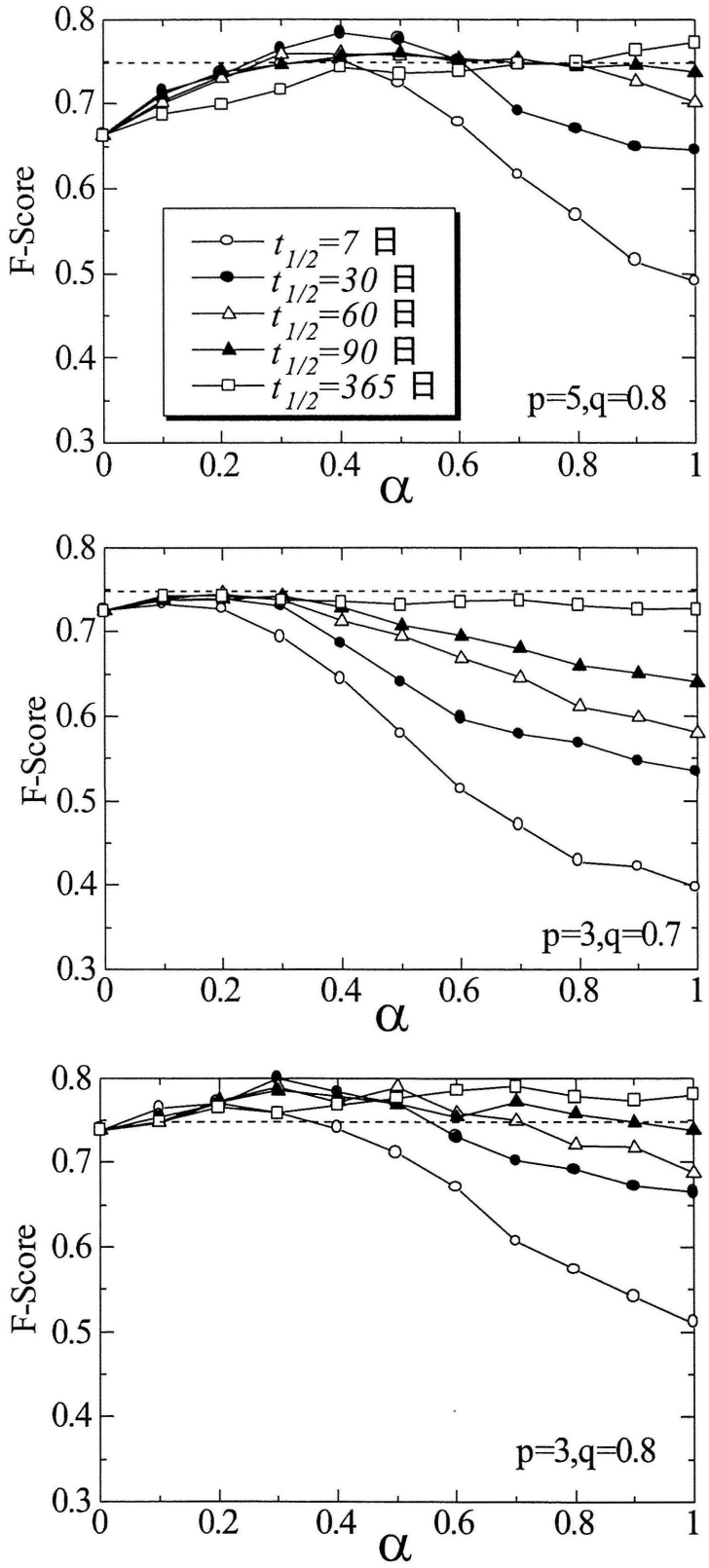

図 4.21 時間類似度に関するパラメータとクラスタリング精度の関係

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，時間を考慮しない比較対象手法からあまり変化しなかった．

また，時間を考慮することによって大きく精度が向上したテストセットとして，“terrorism” 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 (比較対象手法で，トピック抽出の F 値:0.58, F-Score:0.65)．これは，テストセット中に存在するトピックの多くが類似したトピック (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トピックの分類が出来なかったためである．これに対し，時間を考慮することで，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，トピック抽出の F 値:0.71, F-Score:0.75)，時間の考慮がうまく作用した例と言える．

4.5.3 時間類似度利用によるトピック構造の変化

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．

トピック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，クラスタ内のトピックを網羅するような文書であったが，トピックが起こっている最中の記事というよりは，トピック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，時間類似度を導入した手法で抽出されるコア文書は，あるトピックに関連する文書 (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」に関するトピックが挙げられる．このトピックに関しては，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，トピック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ため，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，トピックが盛り上がっている部分の記事を抽出している点が，時間を考慮した場合の特徴的な点であると言える．

4.6 まとめ

本研究では，検索結果全体から特徴的な情報や概要を抽出したいと言う場合に，検索結果に含まれる主要なトピックおよびトピック間の関係を明らかにする事で，ユーザの情報取得を支援しようと言う手法を提案した．

提案手法では，文書間の類似度を基に文書間の関係を示す文書集合グラフを生成し，そのグラフ中のノードの中心性スコアを算出する．そして，これら文書集合グラフとノードの中心性スコアを利用し，各ノードを 4 種のタイプに分類した．これらの情報を用いる事で，「はじめに」で示した問題に対して，以下のような解決策を提供する．

- トピックへのアクセス時
  - 検索結果中の主要なトピックの提示
  - トピック間のつながりやその内容を提示
- 文書へのアクセス時
  - 「トピックの中心的な内容を良く示す文書」や「ノベルティの高いトピックを含む文書」等，文書をトピックの主題との関連性付きで提示

本研究では提案手法に関して，二つの観点から評価を行った．一つ目は基本特性の評価である．新聞記事コーパスを用いて，提案手法をトピック抽出およびクラスタリングのタスクに適用した場合の精度に付いて評価を行い，比較対象手法を上回る精度を示すとともに，パラメータ変化に伴う挙動の変化に付いての知見を得た．また，クラスタリングについては，NTCIR-4 WEB D と同等の評価を行い，3 つのうち，2 つの指標で，NTCIR-4 の最上位システムの値を上回る事を確認し，提案手法の基本特性の高さを確認した．二つ目の評価では，トピック構造マイニングの基本となる文書のタイプ分類が意図通りに機能しているかを評価し，機能していることを確認した．

また，グラフ構造とノードの中心性を利用した文書集合の可視化結果を示し，提案手法で考えるトピック構造が，文書集合の分析に有益であることを示した．

さらに，タイムスタンプ付き文書を解析する場合に，文書集合グラフ構築時に，文書間のタイムスタンプの近さを考慮する手法を提案し，基本特性の評価では，トピック抽出およびクラスタリングの精度が向上することを示した．

第 5 章

結言

本論文では，テキスト情報の検索において，検索結果中に含まれるトピックに注目し，検索結果を構造化する事で，ユーザの情報検索を支援する手法を提案した．以下では，本論文の内容を要約し，今後の展望などについてまとめる．

まず，第 3 章では，キーワードベースの検索システムにおいて，検索条件のあいまいさにより，検索結果の中に所望の情報が埋もれてしまうという問題に対して，検索結果中のトピックに基づいた検索結果のクラスタリングにより検索結果を構造化し，絞り込み検索を支援する手法を提案した．提案手法では，ラベルの候補となるタームの抽出に固有表現抽出を利用し，固有名詞の抽出を行う事を提案した．これは文書のトピックを特徴付ける情報として固有名詞が有益であると考えたためである．また，固有表現抽出によって抽出されたタームに付与される種類 (人名，地名等) をラベルの提示時に利用する事で，同種のラベルをまとめ, 構造化した形で提示できる．二つ目の提案は，上記で抽出したラベル候補からユーザに提示するラベルを選択する新たなラベル選択基準である．この基準では，検索結果内でのタームの重要性と，検索条件との関係性に基づき各タームを評価することで，絞り込み検索に有益であると考えられるラベルの抽出を可能とした．IREX のテストコレクションを用いた評価を行い，提案手法は，比較対象の手法より，検索結果を処理する文書量が少ない場合にも，絞り込み検索に有益なラベルを抽出できる事がわかった．また，提案手法では，比較対象の手法と比較して全体的に高い適合率を示す事がわかった．また，提案手法の二つ目の仮説「検索条件との関連性が高いタームが重要」を表現する式を使い分けることにより，適合率を重視した手法や，適合率と再現率のバランスを重視した手法を使い分け可能であり，適合率を重視した場合には比較対象手法と比較して 15 ポイント以上の高い適合率を示し，バランスを重視した場合にはより少ない検索結果処理量で，比較手法と同等の適合率および再現率を示した．さらに，本手法が，実際のポータル

サービスに応用された事例である「ニュース記事検索システム」,「ブログ記事検索システム」,「話題提示システム」を示すことで，本手法の有効性を示した.

次に，第4章では，特定のキーワードに関係する検索結果全体から特徴的な情報や概要を抽出したいと言う場合に，検索結果中に含まれる主要なトピックおよびトピック間の関係，トピックと個々の文書の関係を明らかにする検索結果の構造化により，ユーザの情報取得を支援する手法を提案した．提案手法では，個々の検索結果文書をノード，文書間の関係をエッジで表現した「文書集合グラフ」により検索結果の集合を表現し，そのグラフ構造とグラフ構造中の各ノードの中心性を利用して，検索結果中のトピック構造を明確化する．基本特性の評価として，ニュース記事を利用したトピック抽出およびクラスタリングの評価を行い，従来のクラスタリングを利用した手法と比較して高い精度を示すことを示し，また，NTCIR-4 WEB D「トピック分類タスク」に基づいたWeb検索結果クラスタリングの評価を行い，3つの評価値のうち2つの評価値で，本タスクの最上位システムを上回る事を示した．また，トピック構造に関する検証では，トピック構造マイニングの基本となる文書のタイプ分類が意図通りに機能しているかを評価し，機能していることを確認した．さらに，ニュース記事やブログ記事のようにタイムスタンプ付きのテキストを処理するために，上記に示した「文書集合グラフ」構築時に，テキスト間のタイムスタンプの近さを考慮する事で，タイムスタンプを意識したトピック構造マイニングを行う手法を提案し，タイムスタンプを利用しない場合と比較し，トピック抽出，クラスタリングの精度が向上することを示した.

「固有表現に注目した検索結果クラスタリング」に関する今後の課題としては，大きく2つの方向性が考えられる．一つは提案手法の適用対象を広くする方向であり，まずは固有名詞だけでなく，他のタームとの組み合せを考慮する事である．もう一つは，よりわかり易いトピックの表現を追求する方向であり，単に固有名詞を提示するだけでなく，固有名詞間のつながりやそのつながりの意味を明確にすることで，よりわかり易い検索が可能になるのではないかと考えている.

また,「グラフ分析を利用したトピック構造マイニング」に関する今後の課題としても大きく二つの方向性が考えられる．一つは，多様な分析を可能とする方向であり，今回は文書間の類似度とタイムスタンプの近さを利用したのみであったが，ハイパーリンクやユーザの文書へのアクセス履歴，文書のメタデータ等を考慮する手法を検討することである．もう一点は，適用対象の大規模化であり，逐次的な情報の更新を考慮することが最初のステップとなると考えている.

謝辞

本研究の遂行ならびに論文の作成にあたり，筑波大学大学院システム情報工学研究科の北川博之教授には懇切なるご指導を賜りました．ここに謹んで感謝の意を表します．また，筑波大学大学院システム情報工学研究科の大保信夫教授，田中二郎教授，西原清一教授，山本幹雄准教授には，本論文の内容についてご指導とご助言を賜わりましたことを深く感謝します．

本研究は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および筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻において行われたものであり，多くの方々のご支援のもと，遂行することができました．

本研究の機会を与えていただいたとともに，ご指導，ご助言をいただいた NTT コミュニケーション科学基礎研究所 外村佳伸氏，NTT サイバーソリューション研究所 奥雅博氏に心から御礼を申し上げます．また，研究の機会を与えていただいたとともに，研究の初期段階から懇切丁寧にご指導をいただいた NTT サイバーソリューション研究所 片岡良治氏に心から御礼を申し上げます．

本研究を進めるうえで，NTT サイバーソリューション研究所 藤村考氏には，様々なご助言をいただいきました．心から感謝致します．また，日常的にご議論，ご助言をいただいた NTT サイバーソリューション研究所の井上孝史氏，植松幸生氏，NTT レゾナント 廣嶋伸章氏 (元 NTT サイバーソリューション研究所) を始めとする先輩同僚の皆様に感謝致します．

第 3 章の技術をシステムとして実現するにあたりご協力いただきました，NTT レゾナント 杉崎正之氏，栗島聡哉氏を始めとする多くの方々に感謝致します．第 4 章の「文書集合グラフ」の可視化において利用した グラフ構造可視化ツールを提供いただきました NTT コミュニケーション科学基礎研究所 斉藤和巳氏に感謝致します．

さらに，これまでの研究遂行においてご指導，ご助力頂いた，筑波大学大学院情報メディア研究科 佐藤哲司教授 (元 NTT コミュニケーション科学基礎研究所)，NTT サイ

バースペース研究所 星隆司氏に，心より感謝致します．また，研究者の先輩として多くの有益なアドバイスを頂きましたNTTコミュニケーション科学基礎研究所 櫻井保志氏に心から感謝致します．

また，北川データ工学研究室に配属以来，名古屋大学情報連携基盤センター 石川佳治教授 (元 筑波大学 助教授)，天笠俊之講師，川島英之講師には御厚意溢れるご助言とご支援を頂きました．ここに厚く御礼申し上げます．また，研究室の先輩として様々なアドバイスやご支援をいただきましたCREST研究員の渡辺陽介氏に心から感謝いたします．また，北川データ工学研究室の諸兄には，日頃より多くのご助言，ご協力をいただき，種々の面でお世話になりました．心から感謝致します．

最後に，本研究を進める上で様々な面で支えていただきました全ての方に感謝の意を表します．

参考文献

[1] Baeza-Yates, R. and Ribeiro-Neto, B.: "Modern Information Retrieval," Addison-Wesley, 1999.

[2] Belkin, N. J.: "Anomalous states of knowledge as a basis for information," Canadian Journal of Information, Vol.5, pp.133–143, 1980.

[3] Bikel, D.M., Schwartz, R. and Weischedel, R.M.: "An Algorithm that Learns Whatś in a Name," Machine Learning, Vol.34, No.1-3, pp.211–231, 1999.

[4] Brin, S., and Page, L., "The anatomy of a large-scale hypertextual Web Search Engine,' Proceedings of the seventh international conference on World Wide Web (WWW7), pp.107–117, Brisbane, Australia, 1998.

[5] Cover, T., and Hart, P.: "Nearest neighbor pattern classification", IEEE Transactions on Information Theory, Volume 13, Issue 1 pp. 21- 27, 1967.

[6] Cui, C. and Kitagawa, H., "Topic Activation Analysis for Document Streams Based on Document Arrival Rate and Relevance," Proceedings of the 2005 ACM symposium on Applied computing (SAC '05), pp.1089–1095, Santa Fe, New Mexico, 2005.

[7] Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W.: "Scatter/Gather: a cluster-based approach to browsing large document collections," Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '92) ,pp. 318–329, Copenhagen, Denmark, 1992.

[8] Dumais, S., and Chen, H.: "Hierarchical classification of Web content," Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '00) , pp.256–263, Athens, Greece, 2000.

[9] Eguchi, K.: "Overview of the Topical Classification Task at NTCIR-4 WEB," Working Notes of 4th NTCIR Workshop Meeting, Vol.Supl. 1, pp.ov-48–ov-55, Tokyo, Japan, 2004.

[10] Erkan, G. and Radev D. R.: "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," Journal of Artificial Intelligence Research, Vo.22, pp.457–479, 2004.

[11] Ferragina, P., and Gulli, A.: "The Anatomy of a Hierarchical Clustering Engine for Web-page, News and Book Snippets," Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM '04), pp. 395–398, Brighton, UK , 2004.

[12] Fujimura, K., Inoue, T., and Sugizaki, M.: "The EigenRumor Algorithm for Ranking Blogs," Proceedings of 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (WWE '05), Chiba, Japan, 2005.

[13] Fukumoto, F., and Suzuki, Y.: "Manipulating large corpora for text classification," Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP '02), pp.196–203, Philadelphia, PA, USA, 2002.

[14] Gallippi, A. F.: "Learning to recognize names across languages," Proceedings of the 16th conference on Computational linguistics (COLING '96), pp. 424–429, Copenhagen, Denmark, 1996.

[15] Grishman, R. and Sundheim, B.: "Message Understanding Conference - 6: A Brief History," Proceedings of the 16th conference on Computational linguistics (COLING '96), pp. 466–471, Copenhagen, Denmark, 1996.

[16] Gulli, A. and Signorini, A.: "The Indexable Web is More than 11.5 Billion Pages," Special interest tracks and posters of the 14th international conference on World Wide Web (WWW '05), pp. 902–903 Chiba, Japan, 2005.

[17] Gusfield, D.: "Algorithms on strings, trees and sequences," Computer Science and Computational Biology, chapter 6, Cambridge University Press, 1997.

[18] Hartigan, J. A., and Wong, M.A.: "A K-means clustering algorithm," Applied Statistics, Vol.28, No. 1, pp.100–108, 1979.

[19] Hayashi, Y., Tomita, J. and Kikui, G.: "Searching text-rich XML documents,"

Proc. of ACM SIGIR 2000 Workshop on XML and Information Retrieval, pp.27–35, Athens, Greece, 2000.

[20] Haveliwala, T., Gionis, A., and Indyk, P.: "Scalable Techniques for Clustering the Web," Proceedings of the 3rd International Workshop on the Web and Databases (WebDB '00), pp. 129–134, Dallas, Texas, USA, 2000.

[21] Hearst, M. A., and Pederson, J. O.: "Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results," Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '96), pp. 76–84, Zurich, Switzerland, 1996.

[22] Hill, D. R.: "A vector clustering technique," In: Samuelson (Ed.), Mechanized Information Storage, Retrieval and Dissemination, North-Holland, Amsterdam, 1968.

[23] Hisamitsu, T., Niwa, Y. and Tsujii, J.: "Measuring Representativeness of Terms," Proceedings of 4th International Workshop on Information Retrieval with Asian Languages (IRAL '99), pp.83–90, Taipei, Taiwan, 1999.

[24] He, X., Ding, C., Zha, H., and Simon, H. D.: "Automatic Topic Identification Using Webpage Clustering," Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM '01), pp.195–202, San Jose, California, USA, 2001.

[25] Isozaki, H.: "Japanese named entity recognition based on a simple rule generator and decision tree learning," Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL '01), pp. 314–321, Toulouse, France, 2002.

[26] Isozaki, H. and Kazawa, H.: "Efficient Support Vector Classifiers for Named Entity Recognition," Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING '02), pp. 390–396, Taipei, Taiwan, 2002.

[27] Kamvar, S. D., Klein, D., and Manning, C. D.: "Spectral Learning," Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI '03), pp.561–566, Acapulco, Mexico, 2003.

[28] Kim, J. D., Ohta, T., Tsuruoka, Y., Tateisi Y. and Collier, N.: "Introduction to the Bio-Entity Recognition Task at JNLPBA." Proceedings of the Joint

Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04), pp.70-75, Geneva, Switzerland, 2004.

[29] Kimura, R., Oyama, S., Toda, H., and Tanaka, K.: "Creating Personal Histories from the Web using Namesake Disambiguation and Event Extraction," Proceedings of 7th International Conference on Web Enginnering (ICWE '07), Como, Italy, 2007. (to appear)

[30] Krishnan, V., and Manning, C. D.: "An effective two-stage model for exploiting non-local dependencies in named entity recognition," Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the the Association for Computational Linguistics (COLING/ACL '06), pp. 1121–1128, Sydney, Australia, 2006.

[31] Kleinburg, J.: "Authoritative sources in hyperlinked environment," Journal of the ACM, Vol.46, No.5. pp.604–632, 1999.

[32] Kumaran, G, and Allan, J.: "Text Classification and Named Entities for New Event Detection," Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04), pp.297–304, Sheffield, United Kingdom, 2006.

[33] Kurland, O., and Lee, L.: "PageRank without hyperlinks: Structural re-ranking using links induced by language models," Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05), pp.306–313, Salvador, Brazil, 2005.

[34] Kurland, O., and Lee, L.: "Respect my authority! HITS without hyperlinks, utilizing cluster-based language models," Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06), pp. 83–90, Seattle, Washington, USA, 2006.

[35] Kummamuru, K., Lotlikar, R., Roy, S., Singal, R., and Krishnapuram, R.: "A hierarchical monothetic document clustering algorithm for summarization and browsing search results," Proceedings of the 13th international conference on World Wide Web (WWW '04), pp. 658-665, New York, NY, USA, 2004.

[36] Kwon, O. W., and Lee, J. H.: "Text categorization based on k-nearest neighbor approach for web site classification," Information Processing and Management,

Vol. 39, Issue 1, pp.25–44, 2003.

[37] Lang, C., and Nguyen, H. S.: "A Tolerance Rough Set Approach to Clustering Web Search Results," Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD '04), pp.515–517, Pisa, Italy, 2004.

[38] Lee, S., and Lee, G. G.: "SiteQ/J: A Question Answering System for Japanese," Working Notes of 3rd NTCIR Workshop Meeting—Part IV: Question Answering Challenge (QAC1), pp. 31–38, Tokyo, Japan, 2002.

[39] Leuski, A.: "Evaluating document clustering for interactive information retrieval," Proceedings of the tenth international conference on Information and knowledge management (CIKM '01), pp. 33–40, Atlanta, Georgia, USA, 2001.

[40] Liu, T. Y., Wan, H., Qin, T., Chen, Z., Ren, Y., and Ma, W. I.: "Site abstraction for rare category classification in large-scale web directory," Special interest tracks and posters of the 14th international conference on World Wide Web (WWW '05), pp. 1108–1109, Chiba, Japan, 2005.

[41] McCallum, A. and Nigam K.: "A Comparison of Event Models for Naive Bayes Text Classification," Proceedings of AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48, Madison, Wisconsin, USA, 1998.

[42] Mihalcea, R. and Tarau P.: "TextRank: Bringing Order into Texts," Proceedings of the Conference on Empirical Methods. in Natural Language Processing (EMNLP '04), pp.404–411, Barcelona, Spain, 2004.

[43] Mihalcea, R., Tarau, P., and Figa, E.: "PageRank on Semantic Networks, with application to Word Sense Disambiguation," Proceedings of The 20st International Conference on Computational Linguistics (COLING '04), pp. 1126–es, Switzerland, Geneva, 2004.

[44] Mihalcea, R.: "Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization," Proceedings of the ACL 2004 on Interactive poster and demonstration sessions (ACL '04), pp. 20–es, Barcelona, Spain, 2004.

[45] Minkov, E, Wang, R. C., and Cohen, W. W.: "Extracting personal names from emails: Applying named entity recognition to informal text," Proceedings of the conference on Human Language Technology and Emprical Methods in Nat-

ural Language Processing (HLT/EMNLP '05), pp. 443–450, Vancouver, British Columbia, Canada, 2005.

[46] Newman, M. E. J.: "The structure and function of complex networks," SIAM Review, Vol.45, No.2, pp.167–256, 2003.

[47] Ohta, M., Narita, H. and Ohno, S.: "Overlapping Clustering Method Using Local and Global Importance of Feature Terms at NTCIR-4 Web Task," Working Notes of 4th NTCIR Workshop Meeting, Vol.Supl. 1, pp. 210–217, Tokyo, Japan, 2004.

[48] Okanohara, D., Miyao, Y., Tsuruoka, Y., and Tsujii, J.: "Improving the scalability of semi-Markov conditional random fields for named entity recognition," Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the the Association for Computational Linguistics (COLING/ACL '06), pp. 465–472, Sydney, Australia, 2006.

[49] Osinski, S.: "Improving Quality of Search Results Clustering with Approximate Matrix Factorisations," Proceedings of the 28th European Conference on IR Research (ECIR '06), pp.167–178, London, UK, 2006.

[50] Pasca, M.: "Acquisition of Categorized Named Entities for Web Search." Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04), pp.137-145, Washington, D.C., USA, 2004.

[51] Rousu, J., Saunders, C., Szedmak, S., and Shawe-Taylor, J.: "Learning hierarchical multi-category text classification models," Proceedings of the 22nd international conference on Machine learning (ICML '05), pp.744–751, Bonn, Germany, 2005.

[52] Sakai, H., Ohtake, K. and Masuyama, S.: "A Retrieval Support System By Suggesting Terms to a User," Proceedings of 19th International Conference on Computer Processing of Oriental Languages (ICCPOL '01), pp. 77–80, Seoul, Korea, 2001.

[53] Salton, G. and Yang, C.G.: "On the Specification of Term Values in Automatic Indexing," Journal of Documentation, Vol.29, pp.351–372, 1973.

[54] Sekine, S. and Nobata, C.: "Definition, dictionaries and tagger for Extended Named Entity Hierarchy," Proceedings of the 4th international conference on Language Resources and Evaluation (LREC '04), pp. 1977–1980, Risbon, Portu-

gal, 2004.

[55] Shinzato, K. and Torisawa, K.: "Extracting Hyponyms of Prespecified Hypernyms from Itemizations and Headings in Web Documents," Proceedings of the 20th international conference on Computational Linguistics (COLING'04), pp.938–944, Geneva, Switzerland, 2004.

[56] Shinzato, K. and Torisawa, K.: "A Simple WWW-based Method for Semantic Word Class Acquisition," Proceedings of the Recent Advances in Natural Language Processing (RANLP '05), pp.493–500, Borovets, Bulgaria, 2005.

[57] Siersdorfer, S., Sizov, S., and Weikum, G.: "Goal-oriented methods and meta methods for document classification and their parameter tuning," Proceedings of the thirteenth ACM international conference on Information and knowledge management (CIKM '04), pp.59-68, Washington D.C., USA, 2004.

[58] Srihari, R., and Li, W.: "A question answering system supported by information extraction," Proc. of the sixth conference on Applied natural language processing (ANLP '00), pp. 166-172, Seattle, Washington, 2000.

[59] Sun, A., Lim, E. P., and NG, W. K.: "Web Classification Using Support Vector Machine." Proc. 4th ACM International Workshop on Web Information and Data Management (WIDM '02), pp.96–99, McLean, Virginia, USA, 2002.

[60] Toda, H., and and Kataoka, R.: "A search result clustering method using informatively named entities", Proc. 7th ACM International Workshop on Web Information and Data Management (WIDM '05), pp 81-86, Bremen, Germany, November 2005.

[61] Yamada, T. Saito, K., and Ueda, N., "Cross-Entropy Directed Embedding of Network Data," Proceedings of the 12th International Conference on Machine Learning (ICML '03), pp.832–839, Washington, DC, USA, 2003.

[62] V.Vapnik, "The Nature of Statistical Learning Theory," Springer, 1995.

[63] Wang, Y, and Kitsuregawa, M.: "Link Based Clustering of Web Search Results," Proceedings of the 2nd International Conference on Web-Age Information Management (WAIM '01), pp. 225–236, Xian, China, 2001.

[64] Wang, Y., and Kitsuregawa, M.: "On Combining Link and Contents Information for Web Page Clustering," Proceedings of the 13th International Conference

on Database and Expert Systems Applications (DEXA '02), Aix en Province, France, pp. 902–913, 2002.

[65] Ward, J. H., "Hierarchical grouping to optimize an objective function," Journal of the American Statistical Association, Vol. 58, No, 301, pp. 236–244, March 1963.

[66] Weiner, P.: "Linear pattern matching algorithms," Proceedings of the 14th Annual Symposium on Foundations of Computer Science (FOCS '73), pp. 1-11, Iowa City, Iowa, USA, 1973.

[67] Yang, Y., Pierce, T., and Carbonell, J. G.: "A Study on Retrospective and Online Event Detection," Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '98), pp. 28–36, Melbourne, Australia, 1998.

[68] Yang, Y, Zhang, J., Carbonell, J., and Jin, C.: "Topic-conditioned Novelty Detection," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02), pp.688–693, Edmonton, Alberta, Canada, 2002.

[69] Zamir, O., and Etzioni, O.: "Grouper: A Dynamic Clustering Interface to Web Search Results," Proceeding of the 8th international conference on World Wide Web (WWW8), pp.1361–1374, Toronto, Canada, 1999.

[70] Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y. and Ma, J.: "Learning to cluster web search results," Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04), pp.210–217, Sheffield, United Kingdom, 2006.

[71] Zhao, J. and He, J.: "Learning to Generate Labels for Organizing Search Results from a Domain-Specified Corpus," Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), pp. 390–396, Hong Kong, China, 2004.

[72] Zhang, L., Pan, Y., and Zhang, T.: "Focused named entity recognition using machine learning," Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04), pp.281–288, Sheffield, United Kingdom, 2006.

[73] Zhao, Y., and Karypis, G.: "Evaluation of Hierarchical Clustering Algorithms for Document Datasets," Proceedings of the 11th international conference on Information and knowledge management (CIKM '02), pp.515–524, McLean, Virginia, USA, 2002.

[74] 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.

[75] 上田 修功, 斉藤 和巳: "多重トピックテキストの確率モデル —パラメトリック混合モデル—," 電子情報通信学会論文誌, Vol.J87-D-2, No.3, pp.872–883, 2004.

[76] 北　研二, 津田和彦, 獅子堀正幹: "情報検索アルゴリズム," 共立出版, 2002.

[77] 佐々木裕, 磯崎秀樹, 鈴木潤, 国領弘治, 平尾努, 賀沢秀人, 前田 英作.: "SVM を用いた学習型質問応答システム SAIQA-II," 情報処理学会論文誌, Vol.45, No.2, pp.635–646, 2004.

[78] 新里圭司, 鳥澤健太郎："Web からの単語クラスの簡単な作り方," 言語処理学会第 11 回年次大会, 2005.

[79] 関根　聡, 井佐原均: "IREX プロジェクト概要," IREX ワークショップ予稿集, pp. 1–5, 1999.

[80] 戸田浩之, 中渡瀬秀一, 片岡良治："特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案", 情報処理学会論文誌:データベース, Vol.46, No.SIG13(TOD27), pp. 99–106, 2005.

[81] 戸田浩之, 片岡良治, 北川博之, "固有表現に着目したニュース記事分類手法の提案," 夏のデータベースワークショップ (DBWS '05), 情報処理学会研究報告 Vol.2005, No.67(2005-DBS-137(I)), 電子情報通信学会技術研究報告 Vol.105, No.171, 2005.

[82] 富田準二, 竹野　浩, 菊井玄一郎, 林　良彦, 池田哲夫："グラフモデルの提案とテキスト検索システムへの適用による評価", 情報処理学会論文誌:データベース, Vol.43, No.SIG2(TOD13), pp. 94–107, 2002.

[83] 成田宏和, 太田　学, 片山　薫, 石川　博：Web 文書検索のための非排他的クラスタリング手法の提案, 第 14 回データ工学ワークショップ (DEWS '03), 2003.

[84] 仲川こころ, 高田喜朗, 関浩之: "可変なカテゴリ構造を用いた文書検索支援手法", 情報処理学会論文誌, Vol. 42, No. 10, pp. 2441–2453, 2001.

[85] 山口雅史, 大島裕明, 小山聡, 田中克己："サーチエンジンのクエリログを利用した同位語の発見", DBSJ Letters, Vol.5, No.2, pp. 1–4, 2006.

[86] BLOGRANGER
http://ranger.labs.goo.ne.jp/br1/

[87] Clusty the clustering search engine
http://clusty.com/

[88] Google:
http://www.google.com/

[89] Mooter - Web Search
http://www.mooter.com/

[90] Vivisimo Clustering Engine
http://search.vivisimo.com/

[91] TopicMaster
http://labs.goo.ne.jp/tm/

[92] ODP - Open Directory Project
http://dmoz.org/

[93] Yahoo!
http://www.yahoo.com/

[94] Yahoo! Directory
http://dir.yahoo.com/

研究業績

博士論文に関する論文

査読付き論文誌

- 戸田浩之, 中渡瀬秀一, 片岡良治, "特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案", 情報処理学会論文誌:データベース, Vol.46, No.SIG13(TOD27), pp. 99-106, 2005 年 9 月.
- 戸田浩之, 北川博之, 藤村考, 片岡良治, 奥雅博, "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, Vol. J90-D, No. 2, pp.292-310, 2007 年 2 月.
- *Hiroyuki Toda*, Ryoji Kataoka, and Masahiro Oku, "Search Result Clustering using Informatively Named Entities", International Journal of Human-Computer Interaction. (in press)

査読付き国際会議

- *Hiroyuki Toda*, and Ryoji Kataoka, "A search result clustering method using informatively named entities", Proc. 7th ACM International Workshop on Web Information and Data Management (WIDM 2005), pp 81-86, Bremen, Germany, November 2005.
- Ko Fujimura, *Hiroyuki Toda*, Takafumi Inoue, Nobuaki Hiroshima, Ryoji Kataoka, Masayuki Sugizaki, "BLOGRANGER - A Multi-faceted Blog Search Engine", Proc. 3nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (WWE2006), Edinburgh, UK, May 2006.
- *Hiroyuki Toda*, Ryoji Kataoka, and Hiroyuki Kitagawa, "Topic Structure Min-

ing for Document Sets using Graph-Based Analysis", Proc. 17th International Conference on Database and Expert Systems Applications (DEXA 2006), LNCS 4080, pp 327-337, Krakow, Poland, September 2006.

- *Hiroyuki Toda*, Ko Fujimura, Ryoji Kataoka, and Hiroyuki Kitagawa, "Topic Structure Mining using PageRank without Hyperlinks", Proc. 9th International Conference on Asian Digital Libraries (ICADL 2006), LNCS 4312, pp 151-162, Kyoto, Japan, November 2006.

査読付き国際会議ポスター論文

- *Hiroyuki Toda*, Mitsutoshi Nagahama, Hidekazu Nakawatase, and Ryoji Kataoka, "A Label-Based Navigation Method using Informatively Named Entities", Proc. 1st ECML/PKDD International Workshop on Knowledge Discovery in Data Streams, Pisa, Italy, pp 99-100, September 2004.
- *Hiroyuki Toda*, and Ryoji Kataoka, "A clustering method for news articles retrieval system", Proc. 14th international conference on World Wide Web (Special interest tracks and posters) (WWW2005), Chiba, Japan, pp. 988-989, May 2005.

査読付き国内会議論文

- **戸田浩之**, 井上孝史, 廣嶋伸章, 杉崎正之, 栗島聡哉, 藤村考, 片岡良治, "マルチファセットブログ記事検索システム「BlogRanger」", 第 17 回データ工学ワークショップ (DEWS2006), 2006 年 3 月.
- **戸田浩之**, 北川博之, 藤村考, 片岡良治, "時間的近さを考慮した話題構造マイニング", 第 18 回データ工学ワークショップ (DEWS2007), 2007 年 2 月.

研究会大会発表

- **戸田浩之**, 長浜光俊, 片岡良治, "特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案", 情報処理学会研究報告, 2004-FI-75, pp. 99-106, 2004 年 5 月
- *Hiroyuki Toda*, and Ryoji Kataoka, "Search Result Clustering Method at

NTCIR-5 Web Query Expansion Subtask", Proc. 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access(NTCIR-5), Tokyo, Japan, December 2005.

- 藤村考, 戸田浩之, 井上孝史, 廣嶋伸章, 片岡良治, 杉崎正之, "マルチファセット型ブログ検索システム BLOGRANGER の開発", 電子情報通信学会技術研究報告 Vol.105, No.650, 2006 年 3 月.

講演

- 戸田浩之, "BLOGRANGER の要素技術 (Core-Technologies of BLOGRANGER)", goo オープンカンファレンス Vol.2 「ブログサーチテクノロジー」, 2006 年 4 月
- 戸田浩之, "BLOGRANGER2.0 の主な特徴 (Main Features of BLOGRANGER2.0)", goo オープンカンファレンス Vol.5 「goo のブログ検索テクノロジーとその未来像」, 2006 年 11 月

その他の論文

査読付き国際会議

- *Hiroyuki Toda*, Toshifumi Enomoto, and Tetsuji Satoh, "Goal-Oriented Information Retrieval Using Feedback from Users", Proc. 2nd International Conference on Web-Age Information Management (WAIM 2001), LNCS 2118 pp 72-79, Xian, China, July 2001.
- Rui Kimura, Satoshi Oyama, *Hiroyuki Toda*, and Katsumi Tanaka, "Creating Personal Histories from the Web using Namesake Disambiguation and Event Extraction", Proc. 7th International Conference on Web Enginnering (ICWE 2007), Como, Italy, July 2007. (to appear)

査読付き国際会議ポスター論文

- Yoshihiko Suhara, *Hiroyuki Toda*, Akito Sakurai, "Event mining from the Blogosphere using topic words", Proc. 1st International Conference on Weblogs and Social Media (ICWSM 2007), Boulder, Colorado, U.S.A., March 2007.

査読付き国内会議論文

- **戸田浩之**, 榎本俊文, 佐藤哲司, "ユーザからのフィードバックを利用した目的指向ナビゲーション手法の検討", 第 12 回データ工学ワークショップ (DEWS2001), 2001 年 3 月.
- **戸田浩之**, 日高東潮, 小島明, 片岡良治, 星隆司, "コンテンツの意味的なメタデータと低レベルの特徴量を統合管理可能なコンテンツガイドの一方式", データベースと Web 情報システムに関するシンポジウム (DBWeb2002), 情報処理学会シンポジウムシリーズ Vol.2002, No.19, pp.351-358, 2002 年 12 月.
- 木村塁, **戸田浩之**, 田中克己, "検索結果スニペットのクラスタリングによる同姓同名人物の特定", 第 17 回データ工学ワークショップ (DEWS2006), 2006 年 3 月.
- 数原良彦, **戸田浩之**, 櫻井彰人, "ブログにおけるイベントマイニングのための適切な関連語の抽出", 第 18 回データ工学ワークショップ (DEWS2007), 2007 年 2 月.

研究会大会発表

- **戸田浩之**, 村本達也, 北角智洋, 星隆司, "映像配信における個人化された広告配信方法の検討", 情報処理学会第 63 回全国大会, 4Y-2, 2001 年 9 月
- **戸田浩之**, 田辺弘実, 日高東潮, 星隆司, "映像配信サービスにおける状況適応型検索システムの提案", 情報処理学会研究報告, 2002-DBS-127, pp. 121-128, 2002 年 5 月
- 田辺弘実, **戸田浩之**, 北角智洋, 星 隆司, "利用者の状況に適応したメタデータ検索機構の提案", 情報処理学会研究報告, 2002-DBS-127, pp. 113-119, 2002 年 5 月
- 日高東潮, **戸田浩之**, 小島 明, 片岡良治, 星 隆司, "XML ベースのコンテンツガイドシステム実現について", 情報処理学会研究報告, 2003-DBS-129, pp. 97-104, 2003 年 1 月

- 向井景洋, 戸田浩之, 片岡良治, "ラベル指向情報検索における分類ラベル統合方式の検討", 情報処理学会・電子情報通信学会情報・システムソサイエティ共催 第3回情報科学技術フォーラム (FIT2004), D-006, 2004年9月
- 戸田浩之, 片岡良治, 北川博之, "固有表現に着目したニュース記事分類手法の提案", 夏のデータベースワークショップ (DBWS2005), 情報処理学会研究報告 Vol.2005, No.67(2005-DBS-137(I)), 電子情報通信学会技術研究報告 Vol.105, No.171, 2005年7月.
- 数原良彦, 戸田浩之, 櫻井彰人, "話題語を手がかりとしたブログからのイベントマイニングの検討", 情報処理学会研究報告, 2006-NL-176, pp. 67-73, 2006年11月