

P2P ネットワークにおける 動的クラスタを用いた検索手法

小林 由依^{†1} 渡辺 俊貴^{†1} 神崎 映光^{†1}
義久 智樹^{†2} 原 隆浩^{†1} 西尾 章治郎^{†1}

近年、各端末（ピア）が互いにサービスを提供しあう P2P ネットワークに対する注目が高まっている．P2P ネットワークで情報検索の効率を向上させるために、内容が類似したデータをまとめ、それらを所持するピアをクラスタ化する研究が行われている．これらの研究の多くは、クラスタがあらかじめ設定されており、動的に変更できない．このため、ある話題のもとで関連するデータが複数のカテゴリに分かれてしまい、検索に必要なメッセージ数が増大するという問題が発生する．そこで本論文では、動的クラスタを用いた検索手法を提案する．提案手法では、話題に合わせて動的にクラスタを生成することで、検索時に発生するメッセージ数を削減する．

A Search Method Using Dynamic Cluster in Peer-to-Peer Networks

YUKIE KOBAYASHI,^{†1} TOSHIKI WATANABE,^{†1}
AKIMITSU KANZAKI,^{†1} TOMOKI YOSHIHISA,^{†2}
TAKAHIRO HARA^{†1} and SHOJIRO NISHIO^{†1}

In recent years, there has been increasing interest in peer-to-peer (*P2P*) networks. To improve the efficiency of information retrieval in P2P networks, many methods for categorizing similar data items and clustering peers have been studied. Most of them construct clusters beforehand in a static manner. Therefore, similar data items regarding a certain topic can be classified into multiple categories. This leads to an undesired increase of network traffic for information retrieval. In this paper, we propose a search method based on dynamic clustering. Our method reduces the number of query messages for information retrieval by dynamically constructing clusters according to arising new topics.

1. はじめに

近年、計算機の普及や高性能化にともない、P2P ネットワークが注目されている．P2P ネットワークでは各計算機端末（ピア）が互いにサービスを提供し合うことで、スケーラビリティや耐障害性が確保できる．一方、各ピアが所持しているデータを把握することが難しく、検索の際には、フラッディングなどの無作為な検索方法が用いられることが一般的であり、検索に必要なメッセージ数が増大する⁸⁾．ここでメッセージ数とは、ピア間でクエリを伝播する際の通信回数を指す．

検索に必要なメッセージ数を削減するために、類似したデータをカテゴリと呼ぶ集合として管理し、同じカテゴリのデータを所持するピアをクラスタ化する研究が行われている^{1)–3),5)–7),9),12),13)}．クラスタとは、ピアどうしを論理ネットワーク上で少ないホップ数で接続させた部分ネットワークである．カテゴリはデータで構成されるものであり、クラスタはピアで構成される点異なる．これらの従来手法では検索の際に、目的のデータを所持している可能性が高いクラスタにクエリを転送することで、少ないメッセージ数で所望のデータを発見できる．

しかし、これらの従来手法のほとんどは、クラスタが静的でネットワーク全体であらかじめ設定されており、動的に変更できない．そのため、ある話題のもとで関連するデータが複数のカテゴリに分かれてしまい、検索に必要なメッセージ数が増大するという問題が発生する．たとえば、ニュース検索においてスポーツや経済のカテゴリがあらかじめ設定されている場合を考える．北京五輪という話題が発生すると、その記事は、スポーツのカテゴリだけでなく、経済のカテゴリにも分類される可能性がある．しかし、従来手法では、静的に構築されたスポーツ、経済の2つのカテゴリのクラスタを検索する必要がある．北京五輪のような流行の話題は、五輪開催前や開催中など特定期間中に非常に多く検索されるため、検索メッセージ数を削減することが求められる．

そこで、本論文では、あらかじめカテゴリごとにクラスタ（静的クラスタ）が構築された P2P ネットワークにおいて、流行の話題の変化にともなって動的にクラスタを生成する検

^{†1} 大阪大学大学院情報科学研究科マルチメディア工学専攻

Department of Multimedia Engineering, Graduate School of Information Science and Technology,
Osaka University

^{†2} 大阪大学サイバーメディアセンター

Cybermedia Center, Osaka University

索手法を提案する．たとえば，北京五輪の例では，北京五輪に関する記事へのアクセスが増加した際に，北京五輪に関するデータを所持するピアで構成されるクラスタを生成する．これにより，流行の話題に対する検索に必要なメッセージ数を削減する．また，ある動的クラスタに伝播されるクエリが減少した場合，その動的クラスタを削除することで，動的クラスタを維持するために必要なメッセージ数を抑える．

以下では，2章で関連研究について述べ，3章で本研究における想定環境について説明する．4章で提案手法について説明し，5章でシミュレーション実験による性能評価を述べる．最後に6章で本論文のまとめと今後の課題について述べる．

2. 関連研究

データを分類する方法に関する研究として，閲覧したニュース記事や受信したメールなどを分類する際に，時系列における単語の出現頻度の変化を考慮し，データの分類や話題の検出を行う研究が多く行われている^{4),14)}．たとえば，文献 4) では，ある一定期間内で出現頻度が高い単語を検出し，検出した出現頻度が高い単語が文中に現れているニュース記事やメールを類似したニュースやメールであるとして 1 つのグループに分類する方法を提案している．また，文献 14) では，ある一定期間内で出現頻度が高い単語を基に，ニュース記事から現実社会で注目されている話題を検出し，検出した話題間の関連を考慮しながら，データを分類する方法を提案している．これらの方法は，データを分類する際に，一時的に注目が高まる話題にも適切に対応できる．このように，ニュースなど，様々な話題が発生する環境において類似したデータを分類する際には，新たな話題の発生を検出し，その話題に対応した分類を行うことが有効である．

一方，P2P ネットワークを用いたファイル共有において，ブログなどを共有するアプリケーションを想定した場合も，一時的に注目が集まる話題を検出し，その話題のデータを所持するピアを適宜クラスタリングすることは非常に有効である．しかし，P2P ネットワークでは，集中的にデータのアクセス頻度などを管理することは困難であり，注目されている話題に関する情報をすべてのピアで共有できない．そのため，文献 4)，14) のように，集中的にデータのアクセス頻度などを管理するサーバによって，話題を検出することを想定している手法は P2P ネットワークには直接適用できない．

また，P2P ネットワークにおける検索効率を向上させるために，類似したデータを持つピアを論理ネットワーク上で少ないホップ数で接続する手法が数多く研究されている．たとえば文献 5) では，データに含まれる単語の出現頻度を要素とするベクトル空間を設定し，

このベクトル空間を等分割してカテゴリを定義する．その後，データをカテゴリに分類し，同じカテゴリに分類されたデータを持つピアを少ないホップ数で接続するクラスタを作成する手法が提案されている．また，文献 1) では，データの内容や年代を考慮してカテゴリを定義している．これらの手法では，検索対象となるカテゴリのデータを持つピアにのみクエリを転送すればよいとため，少ないメッセージ数で適切なデータを多く獲得できる．

しかし，従来手法でデータをカテゴリに分類するためには，集中的にデータのアクセス頻度などを管理する方法がなく，また動的に変化する分類法をすべてのピアで共有することは非常に困難であるため，あらかじめカテゴリを定義しておく必要がある．このため，ある話題のもとで関連するデータが複数のカテゴリに分かれてしまい，検索に必要なメッセージ数が増大するという問題が発生する．

3. 想定環境

本章では，本研究における想定環境について説明する．本研究の応用例は，ニュース記事やブログなどの情報を共有する P2P ネットワークを考えている．なお，本研究の提案手法は文書データを対象としているが，XML や EXIF¹⁵⁾ といった画像のメタデータにも適用可能である．以下では，まず想定しているデータとカテゴリの表現方法，ネットワーク構成とピアの参加脱退，そして検索方法について詳細に説明する．

3.1 データ

各データにはそれぞれデータの特徴を表すベクトル（特徴ベクトル）が付与されているものとする．データの特徴を表現する指標として用いるデータの特徴を表す特徴ベクトルの算出方法は， $\text{tf} \times \text{idf}$ ¹⁰⁾ や，文献 11) で提案されている方法など，既存手法によって与えられているものとする．

3.2 カテゴリ

あらかじめネットワーク全体の共通の情報として，一定の規則に従って類似した内容を持つデータ群を 1 つの集合として表す n 個のカテゴリ ($C_1, \dots, C_n$) が設定されているものとする．カテゴリの特徴を表すベクトルをカテゴリベクトルと呼ぶ．すべてのカテゴリのカテゴリベクトルは各ピアにおいて既知とする．これは，ピアがネットワークに参加する際にすでにネットワークに参加しているピアからその情報を受信することで実現できる．

なお，カテゴリの設定方法は特に規定しないが，たとえば，オンラインニュースサイトやディレクトリ構造型検索エンジンで用いられているカテゴリを利用する方法が適用できる．また，カテゴリベクトルの算出方法は，用意したテストデータを手動でそれぞれのカテゴリ

に分類し、各カテゴリに分類されたデータの特徴ベクトルの平均をとる方法が適用できる。

3.3 データの分類

各データは、3.2 節で述べたカテゴリのうち、最も類似しているものに分類される。各データをカテゴリに分類する際には、各データが各カテゴリとどの程度関連しているかを表す指標である類似度を用いる。データとカテゴリとの類似度 t は、コサイン類似度により式 (1) を用いて計算する。

$$t = \frac{\langle V_k, D \rangle}{\|V_k\| \times \|D\|}. \quad (1)$$

式 (1) において、 V_k 、 D はそれぞれカテゴリ C_k のカテゴリベクトル、およびデータの特徴ベクトルを表す。 $\langle V_k, D \rangle$ は V_k と D の内積を表し、 $\|V_k\|$ は V_k のノルムを示している。

3.4 ネットワーク構成

同一のカテゴリに分類されるデータを所持しているピアで構築されるネットワークを静的クラスタ ($N_1, \dots, N_n$) と呼ぶ。提案手法で用いる動的なクラスタについては、後に 4 章で説明する。各静的クラスタには代表ピアが存在し、代表ピアどうしは互いに接続しているものとする。本研究では、代表ピアで構築されているネットワークを代表ピアネットワークと呼ぶ。ネットワーク構成例を図 1 に示す。1 つのピアは複数のデータを所持しており、自身が所持しているデータが分類されるすべてのカテゴリの静的クラスタに参加する。また、ピアは自身が所属している静的クラスタの代表ピアの情報 (IP アドレスなど) を把握しているものとする。既存研究においても同様の想定がされており、上記の想定は一般性がある。

3.5 ピアの参加と脱退

新たにデータを所持するなどによって新たにネットワークに参加するピア (参加ピア) は、以下の手順で該当する静的クラスタに参加する。その際、参加ピアは、すでにネットワークに参加しているピアからすべてのカテゴリのカテゴリベクトルを受信する。最初に、参加ピアは自身が所持するデータの特徴ベクトルと各カテゴリのカテゴリベクトルの類似度 t を式 (1) を用いて計算し、最も類似度 t が高いカテゴリをデータが分類されるカテゴリとして選択する。

その後、参加ピアは、選択されたカテゴリの静的クラスタの代表ピアに、参加を知らせるメッセージを送信する。これは、すでにネットワークに参加しているピアからそのピアが属するクラスタの代表ピアの情報を受信し、この代表ピアから選択されたカテゴリの静的クラスタの代表ピアの情報を受信することで実現できる。参加を知らせるメッセージを受け取った代表ピアは、自身が代表するクラスタ内のピアの中から隣接ピア数が一定値以下のピアを

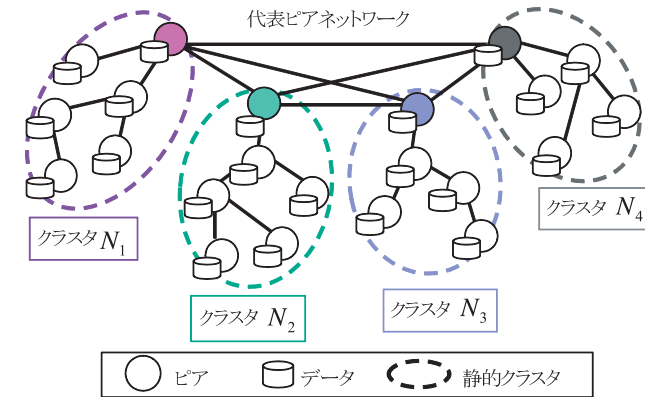


図 1 ネットワークの構成例

Fig. 1 An example of the assumed network.

ランダムに 1 つ選択し、そのピアと参加ピアを接続させる。この隣接ピア数を制限する一定値は、クエリを伝播する際にかかる帯域への負荷と遅延を考慮したうえで、システム管理者が決定する。一定値が大きいほど、帯域への負荷が大きく、遅延は少ない¹⁶⁾。

また、データを削除するなどによってネットワークから脱退するピア (脱退ピア) は、脱退を知らせるメッセージを自身が属するクラスタの代表ピアに送信した後、脱退する。脱退を知らせるメッセージを受け取った代表ピアは、ピアが脱退したことによりクラスタが分断される場合には、脱退ピアの隣接ピアどうしを直接接続させる。なお、ネットワーク障害などによって、ピアが代表ピアに通知することなく脱退してしまった場合でも、そのピアの隣接ピアがクエリを伝播できないことなどにより、脱退を検知できる。

3.6 検索方法

検索の際には、ユーザが入力したキーワードから、そのキーワードの特徴を表すクエリベクトルが生成され、クエリとして送信される。たとえば、図 2 のように、ユーザが「北京五輪」と入力した場合、北京五輪と関連が高い「スポーツ」や「経済」との適合度が高く、それ以外の適合度が低いクエリベクトルが生成される。

ユーザが入力したキーワードからクエリベクトルを生成する具体的な方法は特に規定しないが、たとえば、単語間の関連性が記録されたシソーラスを用いて、入力されたキーワードと、各カテゴリと関連が高い単語との関連の高さを算出し、その値をクエリベクトルの要素とする方法などが適用できる。その際の具体的な例としてシソーラスとして Wordnet¹⁷⁾

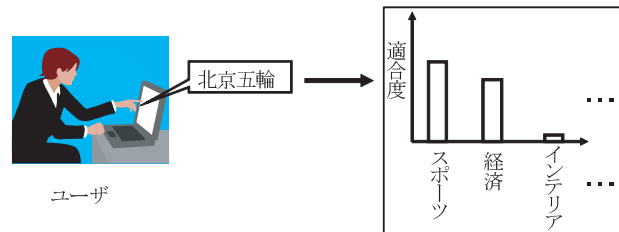


図 2 クエリベクトルの例
Fig. 2 An example of a query vector.

を用いる場合を考える．Wordnet は入力した語に対して意味的な親子関係がある語が出力されるツールであり，このツールを用いて，入力されたキーワードとカテゴリの特徴を表す単語の共通な親を発見するまでに利用した親子関係を数えることで，語の関連度を計測する方法が考えられる．なお，本論文ではこのクエリベクトルは，データの特徴ベクトルと同じ次元のベクトルを想定している．

クエリを発行するピア（クエリ発行ピア）は，クエリベクトルと各カテゴリとの適合度が，式 (2) を用いて算出した α より大きくなるすべてのカテゴリの静的クラスタの代表ピアにクエリを転送し，それらの静的クラスタ内にクエリを伝播させる．

$$\alpha = \frac{\sum_{i=1}^n q_i}{n}. \quad (2)$$

式 (2) において， n は全カテゴリ数， q_i はクエリベクトルのカテゴリ C_i との適合度を表す．クエリ発行ピアは，最初に自身が所属する静的クラスタの代表ピアにクエリを送信する．その際，クエリ発行ピアが複数の静的クラスタに所属する場合には，所属する静的クラスタの代表ピアから 1 つをランダムに選択する．クエリを受信した代表ピアは，クエリの転送先に指定された静的クラスタの代表ピアにクエリを転送する．クエリを受信した各静的クラスタの代表ピアは，自身の受け持つ静的クラスタ内にクエリをフラッディングで転送する．検索時の動作例を図 3 に示す．図 3 において，クエリ発行ピアは，カテゴリ C_1 および C_2 との適合度が高いクエリを発行している．そのクエリはこれらのカテゴリの静的クラスタ N_1, N_2 の代表ピアに転送され，その後，その静的クラスタ内の各ピアにフラッディングで転送される．

クエリを受信したピアは，自身が所持するデータの特徴ベクトルとクエリベクトルの類似度 s を式 (3) を用いて計算する．

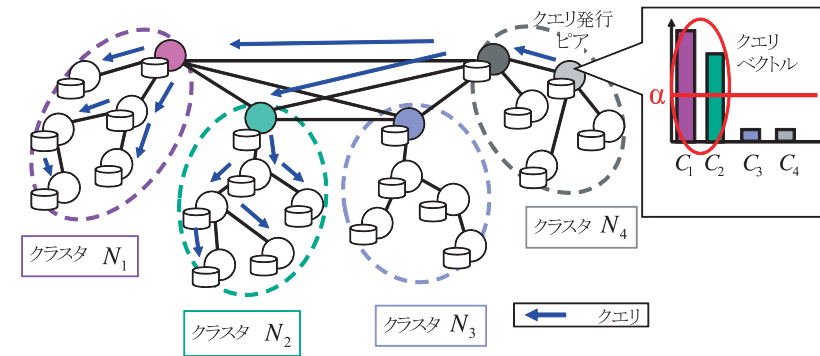


図 3 検索方法
Fig. 3 Search method.

$$s = \frac{\langle Q, D \rangle}{\|Q\| \times \|D\|}. \quad (3)$$

式 (3) において， Q, D はそれぞれクエリベクトル，およびデータの特徴ベクトルを表す． s が定数 β 以上のデータを保持するピアは，そのデータをクエリ発行ピアに返送する．

ユーザは，返送されたデータのうちからいくつかのデータを閲覧して情報を得る．

4. 提案手法

提案手法では，複数のカテゴリに分類されるデータ群へのアクセスの増加を話題の発生と見なし，アクセスの増加を検出して一時的にクラスタを生成する．このクラスタを動的クラスタと呼ぶ．動的クラスタ生成のフローチャートを図 4 に示す．これにより，複数のクラスタに分かれて配置されるデータ群に対して，一時的あるいは定期的にアクセス数が増加した場合に，検索時のメッセージ数を削減する．

本章では，上記の手順，動的クラスタを利用した検索方法，動的クラスタの削除方法，および動的クラスタへの参加と脱退について説明する．

4.1 アクセス増加の検出

複数のカテゴリに分類されるデータ群へのアクセスの増加を検出するために，各静的クラスタの代表ピアは，複数のクラスタに転送されたクエリの情報を記録する．そのため，各クエリ発行ピアは，3.6 節で説明した検索方法においてクエリを送信する際に，複数の静的クラスタへ転送する必要があるクエリには，その旨を示す情報（複数クラスタ転送フラグ）

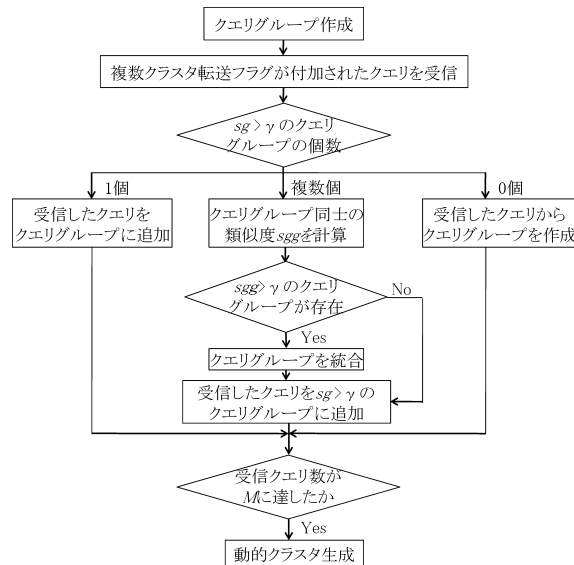


図 4 動的クラスタ生成のフローチャート

Fig. 4 Flowchart of dynamic cluster construction.

を付加する。

各静的クラスタの代表ピアは、受信した複数クラスタ転送フラグが付加されたクエリのうち、類似したクエリベクトルを持つものを図 5 のように 1 つのグループと見なし、その情報を保持する。このグループをクエリグループと呼ぶ。各クエリグループの情報には、グループの特徴を表すための各カテゴリとの適合度から構成されたベクトル（クエリグループベクトル）、受信クエリ数、グループに属するクエリが転送された静的クラスタ（関連クラスタ）、クエリグループの情報が最後に更新された時刻（タイムスタンプ）を含む。たとえば、図 5 のクエリグループ 1 は、過去に類似したクエリを 20 回受け取り、この 20 回のクエリのうち 1 つ以上をカテゴリ C_1 、 C_2 の静的クラスタに転送したことがあることを表している。また、このクエリグループ 1 は、最後に時刻 13:00 において、類似したクエリを受信し、クエリグループの情報を更新している。

4.1.1 クエリグループの作成と更新

複数クラスタ転送フラグが付加されたクエリを受信した代表ピアは、自身の保持する各ク

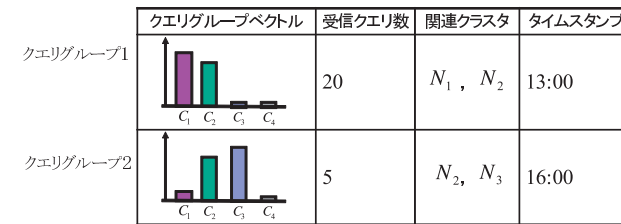


図 5 クエリグループ情報

Fig. 5 Query group information.

エリグループのクエリグループベクトルとクエリベクトルとの類似度を式 (4) を用いて計算する。

$$sg_j = \frac{\langle Q, G_j \rangle}{\|Q\| \times \|G_j\|}. \quad (4)$$

式 (4) において Q はクエリベクトル、 G_j はクエリグループ j のクエリグループベクトルを表す。 G_j の各要素 $g_{j,k}$ ($k = 1, \dots, n$) はクエリグループ j とカテゴリ C_k との適合度を表す。

sg_j がシステムが設定した定数 γ 以上の場合、そのクエリグループベクトルとクエリベクトルは、類似した内容のデータを表していると見なす。類似しているクエリグループの数に応じて、以下の 3 つの場合に分けられる。まず、新しく受信したクエリに対して、類似したクエリグループが 1 つのみ存在する場合は、そのクエリグループに受信したクエリを加えて情報の更新を行う。次に、類似したクエリグループが複数存在する場合には、それらのクエリグループの類似度を確認し、クエリグループどうしが類似していればそれらのクエリグループを統合したうえで、類似したクエリグループに受信したクエリを加えて情報の更新を行う。最後に、類似したクエリグループが 1 つも存在しない場合には、新たにクエリグループを作成する。以下に、それぞれの場合におけるシステムの動作について述べる。

・類似度 sg が γ 以上のクエリグループが 1 つ存在する場合

クエリを受信した代表ピアは、受信したクエリと類似したデータ群を表すクエリグループがすでに 1 つ存在していると判断し、受信したクエリをそのクエリグループに加える。具体的には、そのクエリグループの受信クエリ数の値を 1 増加させ、タイムスタンプを現在の時刻に更新する。また、新たな関連クラスタが存在する場合には、そのクラスタの情報を追加する。さらに、クエリグループベクトル G_j の各要素 $g_{j,k}$ を式 (5) を用いて更新する。

$$g_{j,k} = \frac{g_{j,k}^* \times r_j + q_k}{r_j + 1}. \quad (k = 1, \dots, n) \quad (5)$$

式 (5) において $g_{j,k}^*$ は更新前の $g_{j,k}$ の値である． q_k は Q のカテゴリ C_k との適合度， r_j はクエリグループ j の受信クエリ数を表している．したがって， $g_{j,k}$ は，クエリグループ内の全受信クエリにおけるカテゴリ C_k との適合度の平均値となり，アクセスが高まっているデータ群のカテゴリ C_k との適合度を表している．

・類似度 sg が γ 以上のクエリグループが 2 つ以上存在する場合

類似度が γ 以上のクエリグループどうしは互いに類似したデータ群に対するクエリを表している可能性が高い．そこで，クエリを受信した代表ピアは，これらのクエリグループどうしの類似度 sgg を式 (6) を用いて計算する．

$$sgg = \frac{\langle G_j, G_h \rangle}{||G_j|| \times ||G_h||}. \quad (6)$$

式 (6) において G_j, G_h はそれぞれクエリグループ j, h のクエリグループベクトルを表す．代表ピアは，この値が γ 以上となるクエリグループを 1 つに統合する．統合後のクエリグループのクエリグループベクトル G' の各要素 $g'_{j,k}$ は，式 (7) を用いて設定する．

$$g'_{j,k} = \frac{g_{j,k} \times r_j + g_{h,k} \times r_h}{r_j + r_h}. \quad (k = 1, \dots, n) \quad (7)$$

なお，統合後のクエリグループの受信クエリ数は，クエリグループ j, h それぞれの受信クエリ数の加算値，関連クラスタは， j, h の関連クラスタの和集合，タイムスタンプは現在の時刻に設定する．

クエリグループを統合した後，再び受信したクエリベクトルと統合したクエリグループベクトルとの類似度を計算し，類似度 sg が γ 以上であれば，受信したクエリの情報をそのクエリグループに加えて更新する．

・類似度 sg が γ 以上のクエリグループが 1 つも存在しない場合

クエリを受信した代表ピアは，受信したクエリと類似したデータ群に対するクエリグループが存在しないと判断し，受信したクエリの情報を基に新たなクエリグループを作成する．具体的には，受信したクエリのクエリベクトルをクエリグループベクトルとし，受信クエリ数を 1，タイムスタンプを現在の時刻にする．また，関連クラスタは，そのクエリが転送された静的クラスタとする．

4.1.2 クエリグループの削除

一定期間類似したクエリを受け取らず，更新されなかったクエリグループは，アクセス数

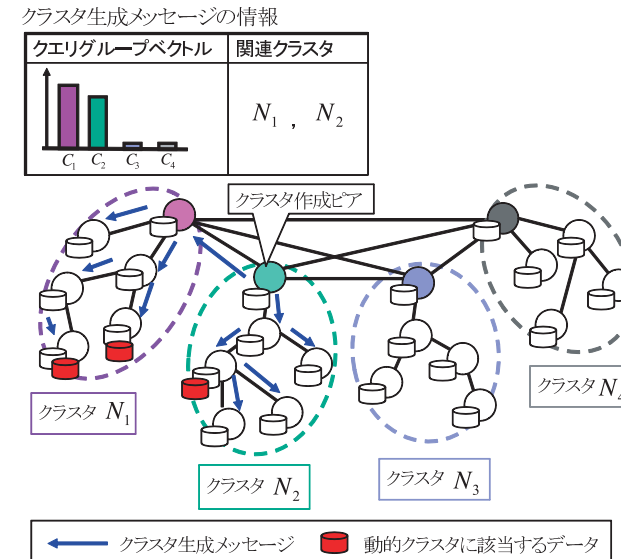


図 6 動的クラスタに含まれるデータの検索

Fig. 6 Searching data included in a dynamic cluster.

が少なく，動的クラスタを生成する必要がないと判断し，代表ピアはその情報を削除する．具体的には，最後に更新された時刻であるタイムスタンプから一定時間 T 以上経過しているクエリグループが存在する場合，そのクエリグループの情報を削除する．なお，クエリグループを保存する期間 T を設定を短くすることで，アクセス頻度が大きい話題にのみ動的クラスタを作成する場合にも対応でき，また，クエリグループを保存する期間 T を設定を長くすることで，アクセス頻度は小さいが長期間アクセスが持続する話題にも動的クラスタを作成することが可能である．

4.2 動的クラスタの生成

各静的クラスタの代表ピアは，1 つのクエリグループに含まれる受信クエリ数が一定数 M に達した場合，そのクエリグループへのアクセスが増加したと判断し，動的クラスタを生成する．

動的クラスタを作成することを決定した代表ピア（クラスタ作成ピア）は，図 6 のように動的クラスタを作成するクエリグループのクエリグループベクトルを付加したメッセージ

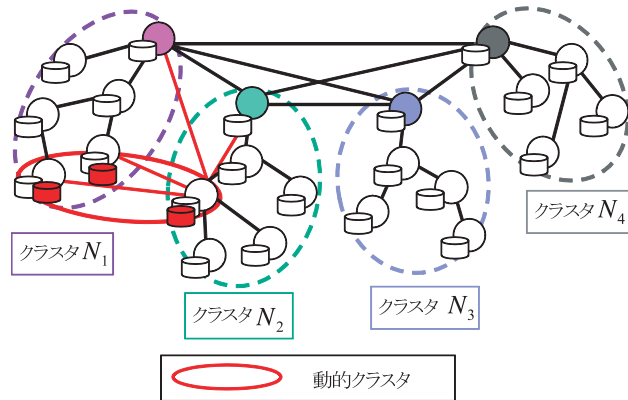


図 7 動的クラスタの構築

Fig. 7 Construction of a dynamic category network.

(クラスタ生成メッセージ)を、自身の静的クラスタ内に伝播させる。また同時に、そのクエリグループの関連クラスタの代表ピアにもクラスタ生成メッセージを送信し、その静的クラスタ内にも伝播させる。クラスタ生成メッセージを受信したピアは、通常の検索と同様、式 (3) を用いて、クラスタ生成メッセージに付加されたベクトルと自身が所持するデータの特徴ベクトルとの類似度を計算する。類似度が定数 δ 以上であれば、そのデータは動的クラスタに含まれると判断し、ピアは自身の ID をクラスタ作成ピアに返送する。

クラスタ作成ピアは、動的クラスタに参加するピアの ID を収集し、ID を返送したピアから順に 3.2 節で述べた参加の手順と同様の方法を用いて、動的クラスタに参加するピアの隣接関係を設定する。その際、動的クラスタに参加するピアから代表ピアを選出し、すべての関連クラスタの代表ピアと接続させる。その後、設定したピアの隣接関係に関する情報に基づいて、クラスタ生成ピアから動的クラスタにおける隣接ピアに隣接関係に関する情報を伝播しながら、互いに接続していくことによって、図 7 のように動的クラスタを生成する。

動的クラスタの情報は、すべての関連クラスタの静的クラスタの代表ピアが保持する。この動的クラスタの情報は、動的クラスタを生成したクエリグループのクエリグループベクトル、および関連クラスタで構成される。

4.3 動的クラスタを利用した検索

各静的クラスタの代表ピアは、クエリを受信した際、そのクエリベクトルと自身が所持する動的クラスタのクエリグループベクトルが類似している場合、その動的クラスタにのみク

関連クラスタの代表ピアが保持している動的クラスタの情報

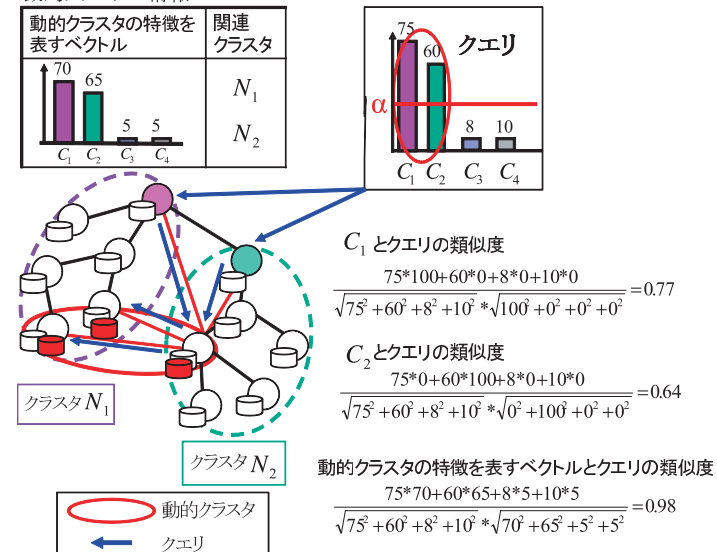


図 8 動的クラスタを利用した検索

Fig. 8 Search method using dynamic cluster.

エリを転送することで、検索時のメッセージを削減する。

具体的には、静的クラスタの代表ピアは、複数クラスタ転送フラグが付加されたクエリを受信した際に、自身が所持する動的クラスタのクエリグループベクトルとクエリベクトルとの類似度を式 (4) を用いて計算する。この類似度が自身が代表する静的クラスタのカテゴリベクトルとクエリベクトルとの類似度より高い場合は、対応する動的クラスタの代表ピアにクエリを転送し、動的クラスタ内に伝播させる。たとえば図 8 では、クラスタ N_1 および N_2 の代表ピアは、クエリを受信した際に、自身が代表する静的クラスタとクエリとの類似度より、自身が管理する動的クラスタとクエリとの類似度の方が高いため、その動的クラスタの代表ピアにクエリを転送させる。その後、クエリを受信した動的クラスタの代表ピアは、動的クラスタ内にクエリをフラディングさせる。

このように、クエリと関連度の高いすべての静的クラスタにクエリを伝播させるのではなく、関連度が高い動的クラスタにのみクエリを伝播させることで、少ないメッセージ数でク

エリを満たすデータを多く獲得できる。

4.4 動的クラスタの削除

一定期間、ある動的クラスタに伝播すべきクエリを受信しなかった場合、アクセス数の減少によってその動的クラスタを維持する必要がなくなったと判断し、その動的クラスタを削除する。具体的には、動的クラスタの代表ピアが、その動的クラスタに伝播したクエリを最後に受信した時刻（タイムスタンプ）を保持しておき、その時刻から一定時間 T 以上クエリを受信しなかった場合、動的クラスタの削除を知らせるメッセージを動的クラスタ内に伝播させる。メッセージを受信したピアは、その動的クラスタ内の隣接ピアとのリンクを切断する。また同時に、動的クラスタの代表ピアは、関連クラスタの代表ピアにもそのメッセージを送信し、関連クラスタの代表ピアはその動的クラスタの情報を削除する。

これにより、不要な動的クラスタが存在することによるオーバーヘッドはそれほど小さくなく、むしろ作成にかかるオーバーヘッドが大きいと考えられるため、確実に不要な動的クラスタのみを削除するようにした。なお、偶然にアクセスの空白期間が生じた場合にも、提案手法では動的クラスタを削除するが、動的クラスタを削除するためまでの期間を長く設定することで対応可能である。

4.5 動的クラスタへのピアの参加と脱退

動的クラスタが作成されている期間中、新たにネットワークに参加するピアは、3.2 節で述べた方法で静的クラスタに参加する際に、関連クラスタとしてその静的クラスタの代表ピアが保持する動的クラスタの情報を受信する。その後、4.2 節で述べた動的クラスタ生成の手順と同様に、動的クラスタのクエリグループベクトルと自身が所持するデータの特徴ベクトルとの類似度を計算する。類似度が定数 δ 以上であるデータを所持している場合は、3.2 節で述べた静的クラスタへ参加する手順と同様に、その動的クラスタに参加する。

また、動的クラスタから脱退するピアは、3.2 節で述べた静的クラスタから脱退する際と同様に、脱退を知らせるメッセージを自身が属する動的クラスタの代表ピアに送信した後、脱退する。脱退を知らせるメッセージを受け取った代表ピアは、ピアが脱退したことによりクラスタが分断される場合には、脱退ピアの隣接ピアどうしを直接接続させることで分断を防ぐ。

5. 性能評価

本章では、提案手法の有効性を検証するために行ったシミュレーション実験の結果を示す。実験では、動的クラスタを生成することによる検索に必要なメッセージ数の削減を明確に評

価することを目的としたため、ピアの参加と脱退については考慮していない。なお、ピアの参加と脱退を考慮したとしても、参加や脱退の際に発行されるメッセージ数は動的クラスタを生成することによって削減されるメッセージ数と比べて十分小さいため、評価結果への影響も十分小さいと考えられる。また、本評価では、静的クラスタのみの P2P ネットワークにおける検索方法との性能比較を行った。

5.1 シミュレーション環境

シミュレーション実験では、ニュース記事を共有する環境をアプリケーション例として想定し、主要なニュースサイトを縮小してモデル化し、以下のようにシミュレーションのパラメータを設定した。カテゴリ数を 20、ネットワーク上に存在するデータ数を 7,500 個と設定し、そのうち 5,000 個のデータは、1 つのカテゴリに対してのみ適合度が高いものとした。具体的には、各データの特徴ベクトルに関して、ランダムに選択した 1 つのカテゴリとの適合度を $[0, 100]$ の範囲でランダムに設定し、それ以外のカテゴリとの適合度を 0 とした。また、残りの 2,500 個は、複数のカテゴリとの適合度が高いデータとした。具体的には類似した特徴ベクトルを持つ 50 個のデータを 1 種類のデータ群とし、50 種類のデータ群を作成した。各データ群の特徴ベクトルは以下の手順で生成した。

- (1) 適合度の高いカテゴリとして、2~5 個のカテゴリをランダムに選択し、それらのカテゴリとの適合度の値を $[30, 80]$ の範囲でランダムに設定する。
- (2) 手順 (1) で選択されたカテゴリの適合度の値に $[-20, 20]$ の範囲のランダムな値を加えて特徴ベクトルを作成する。

手順 (2) の操作を 50 回繰り返すことで、1 つのデータ群の 50 個のデータの特徴ベクトルを作成する。

ネットワークには、500 個のピアが参加し、各ピアはそれぞれ 15 個のデータを持つものとした。各ピアは、各データが分類されるカテゴリの静的クラスタに参加する。ここで 1 つの静的クラスタ内における各ピアの最大隣接ピア数を 3 とし、等しい静的クラスタに参加するピアどうしは、3.2 節で述べた参加の手順と同様の方法で接続するものとした。

1 タイムスロットごとに、ランダムに選択された 1 つのピアがクエリを発行するものとした。発行するクエリのクエリベクトルは、ランダムに選択した 1 つのカテゴリとの適合度を $[30, 80]$ の範囲で設定するものとした。ただし、本実験では、複数のカテゴリに分類されるデータ群へのアクセスが一時的に集中する環境を実現するため、定期的 (250 タイムスロットごと) に複数のカテゴリに分類されるデータ群へのアクセスを増加させた。具体的には、上述した複数のカテゴリに分類される 50 種のデータ群のうちの 1 種のデータ群が頻繁にア

表 1 パラメータ設定
Table 1 Parameter setting.

パラメータ	意味	値
β	クエリに該当する該当データの類似度の閾値	0.85
γ	クエリに該当するクエリグループベクトルの類似度の閾値	0.85
δ	動的クラスタに該当するデータの類似度の閾値	0.75
M	動的クラスタを生成するクエリ数の閾値 (個)	20
T	クエリグループを削除するまでのタイムスロット	70

クセスされる期間を、各データ群ごとにそれぞれ 150 タイムスロットずつ設けた。複数のカテゴリに分類されるデータ群へのアクセスが増加している期間中に発行するクエリは、上述したデータの作成手順 (1) で作成した特徴ベクトルの各カテゴリの適合度の値に、 $[-20, 20]$ の範囲のランダムな値を加えることで作成した。なお、本実験では、全シミュレーションを 15,000 タイムスロットとしており、各データ群へのアクセス集中時 (150 タイムスロット) はそれに比べて大幅に短いものとしている。これは、複数のカテゴリに分類されるデータ群へのアクセスの集中は、一時的なものであるという想定に基づいている。その他の閾値に関する各種のパラメータは、各評価で明記していない限り、表 1 に示すように設定した。

また、実際に返送されたデータの中で、検索要求に該当するデータの割合を表す適合率は、すべての検索において 1 となる。これは、本研究のシミュレーション実験では、実際に返送されたデータはすべてクエリに該当すると設定したためである。

5.2 評価と考察

5.2.1 タイムスロットにおける評価

まず、提案手法および比較手法においてタイムスロットごとの以下の項目を評価した。

● 総メッセージ数

検索時に発行されるメッセージ数、および、提案手法において動的クラスタを生成、削除する際に必要なメッセージ数の合計。ただし、シミュレーション開始時からの累計とする。

● 再現率

各クエリにおけるネットワーク上に存在する該当データのうち、実際に返送されたデータの割合。値が大きいほど、要求したデータを多く獲得できていることを示す。

結果を図 9、図 10 に示す。図 9 (b) は、可視性を高めるため、図 9 (a) においてタイムスロットが 1~200 の部分を抜粋している。また、図 9 において、比較手法では動的クラスタを生成せず、動的クラスタの生成、削除に必要なメッセージは存在しないため、提案手法

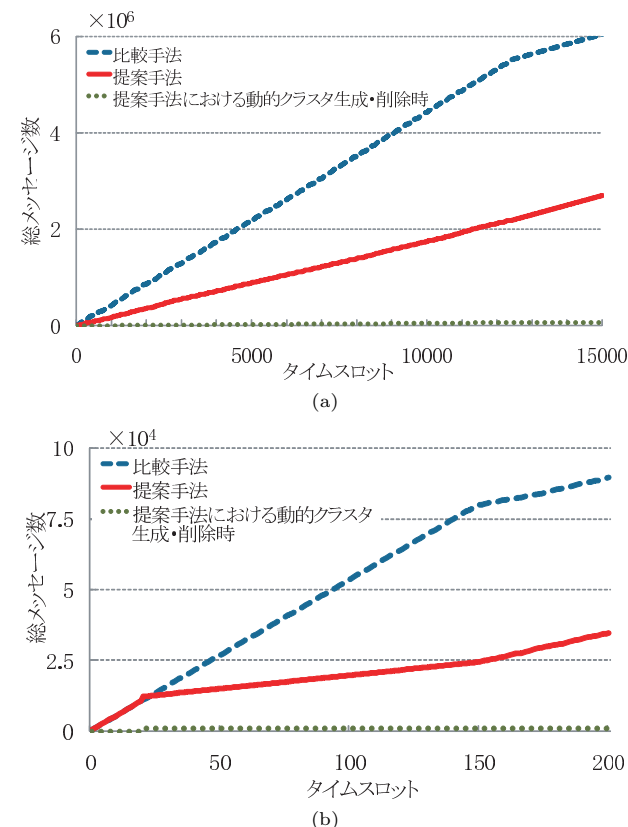


図 9 総メッセージ数
Fig. 9 The number of total messages.

における動的クラスタ生成、削除時に発行されているメッセージの総数のみ示す。これらの図において、横軸はタイムスロットを表し、縦軸はそれぞれ総メッセージ数、再現率を表している。シミュレーション実験では、各タイムスロットごとに 1 つのピアがクエリを発行して検索を行う。250*x ~ 250*x + 150 [タイムスロット] (x は 0 以上の整数) の期間には、複数のカテゴリに分類される話題が発生し、その話題に関するデータ群への検索が集中的に発生するものとした。上記以外の期間には、1 つのカテゴリに分類できる話題に関する検

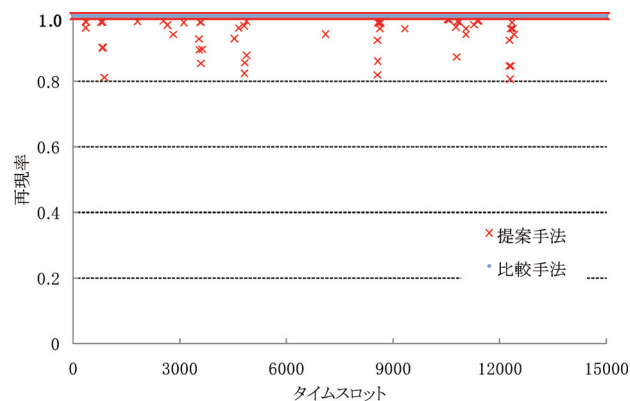


図 10 再現率
Fig. 10 Recall.

索が発生するものとした。

図 9 の結果より、提案手法では、比較手法と比較して、総メッセージ数を大幅に削減できることが分かる。これは、比較手法では、複数のカテゴリとの適合度が高いデータ群へのクエリは適合度の高いすべてのカテゴリの静的クラスタに伝播される一方で、提案手法では、それらのデータ群を所持するピアからなる動的クラスタを構築することで、必要なピアにのみクエリを伝播できるためである。たとえば本実験では、タイムスロットが 0 から 150 の間、複数のカテゴリに分類されるデータ群へのアクセスが集中しており、比較手法と比べて提案手法におけるメッセージ数の増加量が抑えられている。また、提案手法では、動的クラスタの生成、削除のためにメッセージが発生するが、この量はきわめて小さいことが分かる。

また、図 10 の結果より、提案手法では、必要なデータの大部分を獲得できることが分かる。しかし一方で、再現率が 1 にならない検索が存在する。提案手法では、類似したクエリを 1 つのクエリグループにまとめる際に、各クエリのクエリベクトルを平均化してクエリグループベクトルを算出している。この場合、クエリグループに含まれる個々のクエリのクエリベクトルとそのクエリグループベクトルに違いが生じるため、個々のクエリに対して該当するすべてのデータがその動的クラスタ内に配置されるとは限らない。そのため、動的クラスタにのみクエリをフラッシングさせても必要なデータを獲得できない場合があり、再現率が低下している。しかし、提案手法において再現率が低下した検索は、全体の 0.6%程度であり、性能の低下は非常に小さいものといえる。また、再現率が低下した場合でも、ク

エリを満たす全データの 8 割以上が獲得できている。本論文で想定するニュースやブログなどの記事を考慮した場合、ユーザが検索結果となる全データを閲覧することは稀であるものと考えられる。そのため、たとえば、200 人の閲覧者のうち、199 人が要求した全データを獲得でき、残り 1 人についても 8 割のデータが入手できる環境は、十分に実用に耐えるものと考えられる。

5.2.2 閾値 δ の影響

提案手法では、動的クラスタに該当するデータの類似度の閾値 δ の値を大きくすることにより、1 つの動的クラスタに該当するデータの数が少なくなり、検索メッセージ数や検索成功率に大きな影響を与える。そこで、提案手法における閾値 δ の影響を調べるために、 δ を 0.5 から 0.8 の間で 0.05 ずつ変化させた場合の、以下の項目を評価した。実環境において、検索システム管理者は同様の評価を用いて、パラメータ δ の値を決定する。また、提案手法では、 δ の値を変化させても、静的クラスタを対象とした検索時のメッセージ数や返信データの数は変化しない。そのため、複数のカテゴリに分類されるデータ群を対象とした検索における以下の項目を評価した。

- 平均検索メッセージ数
複数のカテゴリに分類されるデータ群への検索 1 回ごとに発行されるメッセージ数の平均値
- 平均再現率
複数のカテゴリに分類されるデータ群への検索 1 回ごとの、ネットワークに存在する検索要求に該当するデータの中で実際に返送されたデータの割合の平均値

平均検索メッセージ数と平均再現率の結果をそれぞれ図 11 および図 12 に示す。これらの図において横軸はパラメータ δ を表し、縦軸はそれぞれ平均検索メッセージ数、平均再現率を表す。提案手法ではパラメータ δ の値が大きいくほど、動的カテゴリのクエリグループベクトルとその動的クラスタに含まれるデータの特徴ベクトルの類似度の条件が厳しくなる。

図 11 の結果より、パラメータ δ が大きくなるにつれて、平均検索メッセージ数が減少していることが分かる。これは、パラメータ δ を大きくすることにより、動的クラスタに配置されるデータが厳選され、その動的クラスタに参加するピアの数が少なくなるためである。これにより、検索時のメッセージ数が削減される。

一方、図 12 の結果より、パラメータ δ が大きくなるにつれて、平均再現率が減少している。これは、パラメータ δ を大きくし、動的クラスタに配置されるデータの条件を厳しくすると、各クエリに該当する必要なデータが動的クラスタ内に配置されない可能性が高く

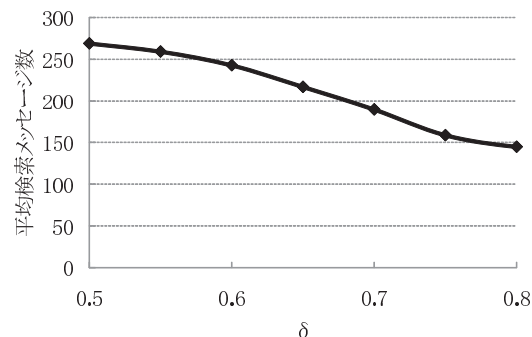


図 11 平均検索メッセージ数

Fig. 11 Average number of query messages.

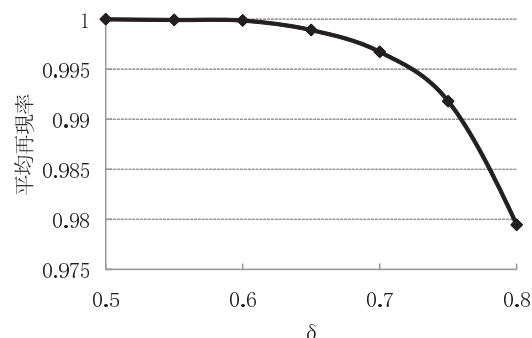


図 12 平均再現率

Fig. 12 Average recall.

なるためである．したがって，図 11，図 12 の結果より，動的クラスタを生成する際には，検索に必要なメッセージ数と再現率のトレードオフを考慮した適切な δ の値を設定する必要がある．

6. おわりに

本論文では，あらかじめカテゴリごとにクラスタが構築された P2P ネットワークにおいて動的にクラスタを生成する検索手法を提案した．提案手法では，複数のカテゴリに分類さ

れるデータ群に対するアクセス数の増加を検知し，一時的にそれらのデータ群に適する新たなクラスタを生成することにより，検索に必要なメッセージ数を削減する．シミュレーション実験により，提案手法によって複数のクラスタに配置されるデータ群に対するアクセス数が増加した場合でも，少ない検索メッセージ数で必要なデータを取得できることを確認した．

今後は，データの更新によるデータの特徴ベクトルの変化を考慮し，動的クラスタに配置されている多くのデータの特徴ベクトルが変化した場合に，動的クラスタのクエリグループベクトルを迅速かつ柔軟に更新する手法を検討する．

また，本論文では，動的クラスタの生成による検索に必要なメッセージ数の削減量を明確に表すために，ピアの参加や脱退が発生しない環境での評価を行った．ここで，ピアの参加脱退を考慮したとしても，手法間の優劣に差はなく，全体的に性能が少し低下するのみと考えられる．しかし，その影響の度合いについて詳細に評価することは有意義であるため，今度はピアの参加脱退を考慮したシミュレーション実験を行うことも検討している．

最後に，本論文で性能評価のために行ったシミュレーション実験では，提案手法の動作や特性を調査するために，データ分布およびクエリの発生頻度として，単純なモデルを使用した．今後は，実際の文書データや検索キーワードデータを用いた性能評価を行う予定である．

謝辞 本研究の一部は，特定領域研究（18049050）の研究助成によるものである．ここに記して謝意を表す．

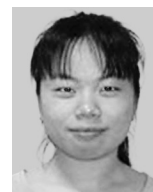
参 考 文 献

- 1) Crespo, A. and Garcia-Molina, H.: Semantic overlay networks for P2P systems, Technical Report, Computer Science Department, Stanford University, pp.23–32 (2002). doi:10.1007/11574781_1
- 2) Huang, J., Li, X. and Wu, J.: A class-based search in unstructured P2P networks, *Proc. Int. Conf. on Advanced Networking and Applications (AINA 2007)*, pp.76–83 (May 2007). doi:10.1109/AINA.2007.8
- 3) Jin, H., Ning, X., Chen, H. and Yin, Z.: Efficient query routing for information retrieval in semantic overlays, *Proc. Int. Workshop on Selected Areas In Cryptography (SAC 2006)*, pp.1669–1673 (Apr. 2006). doi:10.1145/1141277.1141672
- 4) Kleinberg, J.: Bursty and hierarchical structure in streams, *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2002)*, pp.91–101 (July 2002). doi:10.1145/775047.775061
- 5) Li, M., Lee, W.C. and Sivasubramanian, A.: Semantic small world: An overlay network for peer-to-peer search, *Proc. Int. Conf. on Network Protocol (ICNP 2004)*, pp.228–238 (Oct. 2004). doi:10.1109/ICNP.2004.1348113

- 6) Linari, A. and Weikum, G.: Efficient peer-to-peer semantic overlay networks based on statistical language models, *Proc. Int. Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR'06)*, pp.9–16 (Nov. 2006). doi:10.1145/1183579.1183582
- 7) Lu, J. and Callan, J.: Content-based peer-to-peer network overlay for full-text federated search, *Proc. 8th RIAO Conf. on Large Scale Semantic Access to Content (RIA0'07)* (2007).
- 8) Lv, Q., Cao, P., Cohen, E., Li, K. and Shenker, S.: Search and replication in unstructured peer-to-peer networks, *Proc. Int. Conf. on Supercomputing (ICS 2002)*, pp.84–95 (June 2002). doi:10.1145/514191.514206
- 9) Norvag, K., Doukeridis, C. and Vazirgiannis, M.: Semantic overlays for P2P web searching, Technical Report, AUEB (2005).
- 10) Salton, G.: Developments in automatic text retrieval, *Science*, Vol.253, pp.974–979 (1991).
- 11) Shirakawa, M., Nakayama, K., Hara, T. and Nishio, S.: Concept vector extraction from wikipedia category network, *Proc. Int. Conf. on Ubiquitous Information Management and Communication (ICUIMC2009)*, pp.71–79 (Jan. 2009). doi:10.1145/1516241.1516255
- 12) Tang, C., Xu, Z. and Dwarkadas, S.: Peer-to-peer information retrieval using self-organizing semantic overlay networks, *Proc. SIGCOMM 2003*, pp.175–186 (Aug. 2003). doi:10.1145/863955.863976
- 13) Tang, C., Xu, Z. and Dwarkadas, S.: On scaling latent semantic indexing for large peer-to-peer systems, *Proc. Int. Conf. on Special Internet Group on Information Retrieval (SIGIR 2004)*, pp.112–121 (July 2004). doi:10.1145/1008992.1009014
- 14) Wang, X. and McCallum, A.: Topics over time: A non markov continuous time model of topical trends, *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2006)*, pp.424–433 (Aug. 2006). doi:10.1145/1150402.1150450
- 15) Yasumoto, R.: EXif Reader—デジタルカメラデータ解析ツール．
入手先 <http://www.rysys.co.jp/exifreader/jp/> (参照 2009-05-13)．
- 16) 後藤嘉宏, 阿多信吾, 村田正幸: P2P サービスにおける物理ネットワークを考慮した論理トポロジー構築手法, 映像情報メディア学会技術報告 (2002).
- 17) 独立行政法人情報通信研究機構: 日本語 Wordnet .
入手先 <http://nlpwww.nict.go.jp/wn-ja/> (参照 2009-09-16) .

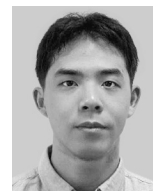
(平成 21 年 5 月 25 日受付)

(平成 21 年 12 月 17 日採録)



小林 由依

2009 年大阪大学工学部電子情報エネルギー工学科卒業．現在，同大学大学院情報科学研究科博士前期課程在学中．P2P ネットワークにおける検索方法に興味を持つ．日本データベース学会の学生会員．



渡辺 俊貴 (学生会員)

2006 年大阪大学工学部電子情報エネルギー工学科卒業．2007 年同大学大学院情報科学研究科博士前期課程修了．現在，同大学院情報科学研究科博士後期課程在学中．P2P ネットワークにおける複製管理に興味を持つ．電子情報通信学会，日本データベース学会の学生会員．



神崎 映光 (正会員)

2002 年大阪大学工学部情報システム工学科卒業．2004 年同大学大学院情報科学研究科博士前期課程修了．2005 年同大学院情報科学研究科博士後期課程中退後，同大学院情報科学研究科マルチメディア工学専攻特任助手を経て，2007 年より同助教となり，現在に至る．博士 (情報科学)．2008 年，2009 年本学会論文賞受賞．2009 年本学会山下研究記念賞受賞．移動体ネットワーク，通信プロトコル，分散処理に興味を持つ．IEEE，電子情報通信学会，日本データベース学会の各会員．



義久 智樹 (正会員)

2002 年大阪大学工学部電子情報エネルギー工学科卒業．2003 年同大学大学院情報科学研究科マルチメディア工学専攻博士前期課程を修了し，2005 年同専攻博士後期課程修了．博士 (情報科学)．2005 年京都大学学術情報メディアセンター助教．2008 年大阪大学サイバーメディアセンター講師を経て 2009 年より同准教授となり，現在に至る．この間，カリフォルニア大学アーバイン校客員研究員．インターネット放送およびセンサネットワークに興味を持つ．電子情報通信学会，IEEE 各会員．



原 隆浩（正会員）

1995 年大阪大学工学部情報システム工学科卒業．1997 年同大学大学院工学研究科博士前期課程修了．同年同大学院工学研究科博士後期課程中退後，同大学院工学研究科情報システム工学専攻助手，2002 年同大学院情報科学研究科マルチメディア工学専攻助手，2004 年より同大学院情報科学研究科マルチメディア工学専攻准教授となり，現在に至る．工学博士．1996 年本学会山下記念研究賞受賞．2000 年電気通信普及財団テレコムシステム技術賞受賞．2003 年本学会研究開発奨励賞受賞．2008 年，2009 年本学会論文賞受賞．データベースシステム，分散処理に興味を持つ．IEEE，ACM，電子情報通信学会，日本データベース学会の各会員．



西尾章治郎（フェロー）

1975 年京都大学工学部数理工学科卒業．1980 年同大学大学院工学研究科博士後期課程修了．工学博士．京都大学工学部助手，大阪大学基礎工学部および情報処理教育センター助教授，大阪大学大学院工学研究科情報システム工学専攻教授を経て，2002 年より大阪大学大学院情報科学研究科マルチメディア工学専攻教授となり，現在に至る．2000 年より大阪大学サイバーメディアセンター長，2003 年より大阪大学大学院情報科学研究科長，その後 2007 年より大阪大学理事・副学長に就任．この間，カナダ・ウォータールー大学，ピクトリア大学客員．データベース，マルチメディアシステムの研究に従事．現在，Data & Knowledge Engineering 等の論文誌編集委員．本会理事を歴任．本会論文賞を受賞．電子情報通信学会フェローを含め，ACM，IEEE 等 8 学会の各会員．