P2P を応用した類似文書検索システム

木浦 正博

1 はじめに

近年，ファイル，動画共有ソフトウェアなどの登場により，Peer-to-Peer(P2P) システムが注目を集めている．P2P システムでは，従来の Client-Server システムと比較して，各ノードが特定の 1 ノードに依存することなく，動作させることができる．そのため，その自律分散性から，様々な自律分散ネットワークへの応用が図れるものと考えられる．

本研究では，このような P2P の技術の様々な分野へ応用例として，P2P のネットワーク技術を文書検索へ用いることを目指す．検索技術においては，各文書間の類似度を求める一手法として Vector Space Model(VSM) が知られている．VSM では，

1. 各文書における複数の語とその出現頻度を特徴ベクトルとして抽出
2. 各文書における特徴ベクトルの内積を求めることにより，各文書間の類似度を算出

という過程を経て，各文書間の類似度を求める．そのため，N 個の文書間における類似度算出には，$O(N^2)$ だけの計算が必要となる．

一方，P2P では，部分的な情報から全体情報を予測する手法として，Vivaldi[1] が知られている．Vivaldi では，ネットワークにおけるノード間の遅延時間を距離関係とし，各ノードを座標空間上に配置することで，計測していない遅延時間を座標間距離から予測する手法である．この Vivaldi を用いて，各文書のクラスタリングを行う．

2 Peer-to-Peer

2.1 Peer-to-Peer の概要

現在我々が利用しているネットワークのほとんどが，組織や団体，国家に依存したネットワークであり，ユーザーはプロキシサーバーや ISP を介してインターネットに接続している．P2P ネットワークはこのような既存のネットワークの上位に仮想ネットワークを構築し，ユーザーがサーバーや組織などを意識することなくネットワークにアクセスするための技術である．そのサービスはインスタントメッセージソフトウェアや近年，よく利用されるファイル共有ソフトウェアなどによって提供されるものが多い．

2.1.1 分散座標ネットワーク系 Vivaldi

Vivaldi は，P2P において物理的なばねの伸縮性を利用し，ネットワーク上のノード間の遅延時間から仮想座標空間にノード群の地図を描く手法である．Vivaldi には以下のような特徴がある．

- 分散型ネットワーク座標系
  各ノードが自律的に自身の座標を決定する．
- ばねの原理
  ユークリッド距離と実測値との誤差を徐々に修正し，最適化する．

Vivaldi ではネットワークに参加しているすべてのノードにおいて，E を推定誤差，L_{ij} をノード ij 間の遅延時間，x_i をノード i の座標とした場合に，(1) 式によって求められる 2 乗誤差を最小にすることが目的である．

$$E = \sum_i \sum_j (L_{ij} - \|x_i - x_j\|)^2 \tag{1}$$

これは物理的なばねの運動においてばね系全体のばねエネルギーを最小化することを模倣している．しかし，ネットワーク全体の推定誤差を求めることは困難であるため，各ノードの座標を分散最適化する．

まず，2 つのノード ij において，ノード i が調節すべき方向ベクトルは単位ベクトルを $u(x_i - x_j)$ として，(2) 式によって調節する値を決める．

$$F_{ij} = (L_{ij} - \|x_i - x_j\|) \times u(x_i - x_j) \tag{2}$$

次に方向ベクトルと現在の座標からノード i は F_i を最小とする方向に移動する．あるノードとその相手ノードとの距離が遅延時間よりも長ければ，徐々に座標を変更し距離を短くとる．逆に，遅延時間が長ければ，距離を長くとるように座標を調節する．しかし，その収束性は通信先のノードによるため座標が発振してしまう可能性がある．BambooDHT[2] では定数 δ を時間の経過とともに減少させることにより，収束をはかる．(3 式)

$$x_i = x_i + F_i \times \delta \tag{3}$$

P2P ネットワークでは，一度にすべてのノードが参加するのではなく，ある程度の間隔をあけてノードが参加するため，あるノードが参加する際には，ネットワーク内のノードの座標は最適化されているものと考える．

2.1.2 Peer-to-Peer システムにおける動的，階層的クラスタリング

P2P システムにおけるクラスタリングは，上田らによって提案されている．[3] この手法では，各ノードを動的かつ階層的にクラスタリングしている．各ノードが保持すべき情報は，クラスタ表，クラスタ情報の 2 種類があり，それぞれのノードが持つこれらの情報を参照し合うことによりクラスタリングが実現される．

このモデルでは，各クラスタ階層におけるクラスタ数を固定している．1 に示すように各ノードは，自身が参加する兄弟クラスタの 1 ノードに対してリンクを持って

おり，自身が参加している末端クラスタにおいては，同一クラスタ内のすべてのノードに対してリンクを持っている．

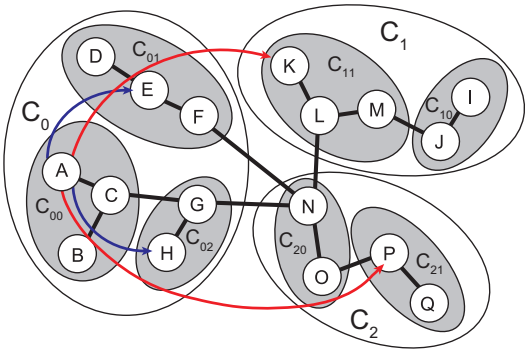

Fig.1　クラスタリングの様子（出典：自作）

- クラスタが保持する情報のデータ構造各クラスタのノードが保持するデータを 1 と 2 にに示す．

Table1　クラスタ表

階層	所属クラスタ	兄弟クラスタ
1	C_0(のクラスタ情報)	C_1
2	C_{00}	C_{01}

- クラスタの形成- ノードの追加
  クラスタに参加するノードは他の P2P システムと同様に，すでにクラスタに参加している 1 ノードを知っているものとする．各クラスタ階層においてクラスタ数の上限値に達していない場合には，新たにクラスタを作成し，上限値に達している場合には，もっともネットワーク的に近いクラスタに参加する．

3　P2P を応用した類似文書検索システム

本研究では，まず，一部の文書間で TF-IDF と VSM を用いて，文書間の距離を算出する．そして，その距離を基に，Vivaldi によって仮想空間上に文書を配置し，P2P 手法による動的，階層的クラスタリングを行う．

3.1　システムの概要

本システムでは，前章で述べた Vivaldi と P2P クラスタリングアルゴリズムを用いて，類似文書検索システムを提案する．このシステムでは，文書間の類似度を距離とし，クラスタを形成する．この類似度を求めるアルゴリズムを 3.2 で述べる．

3.2　TF-IDF と Vector Space Model

文書間の関連度を求める手法には様々な手法があるが，ここでは，TF-IDF と VSM を用いた手法を用いる．TF-IDF は，各単語が特定文書をどれだけ特徴付けているかを表す指標である．特定文書内における単語 i の出現頻度 n_i(Term Frequency: TF) は (4) 式によって表され，

$$tf_i = \frac{n_i}{\sum_k n_k} \tag{4}$$

文書群内における語の出現頻度 (Inverse Document Frequency: IDF) は総ドキュメント数 $|D|$ と単語 i を含むドキュメント数 d_i によって (5) 式のように表される．

$$idf_i = log\frac{|D|}{|d_i|} \tag{5}$$

以上から，TF-IDF は (6) 式のように表される．

$$tfidf_i = tf_i \times idf_i \tag{6}$$

2 つの文書 A,B において出現するすべての単語の TF-IDF をベクトルとすれば，文書間の類似度は (7) 式で表される．

$$sim(A,B) = \frac{\vec{V}(A)\cdot\vec{V}(B)}{|\vec{V}(A)||\vec{V}(B)|} \tag{7}$$

この類似度を文書間の距離として，Vivaldi によって仮想座標空間にマッピングを行う．

3.3　システムの動作

これまで述べたことから，クラスタリングのためのシステムの動作は以下のようになる．

1. システムへの文書の追加
2. 文書と他のいくつかの文書間で類似度を算出する
3. Vivaldi とクラスタリングアルゴリズムによってクラスタリングを行う
4. 以後，システムへ文書が追加される度にこの動作を繰り返す

4　今後の課題

本研究における今後の課題を以下に示す．

- システムの実装
  TF-IDF と VSM によって類似度を求め，Vivaldi によって文書を仮想仮想座標空間上にマッピングする処理までは現在完成している．今後は，P2P によるクラスタリングアルゴリズムを実装する．
- 実験
  本研究室が管理する ISDL Report システム [4] や Wikipedia[5] のデータを用いてクラスタリング実験を行う．

5　まとめ

本研究では，様々な分野への応用が可能であろうと考えられる，P2P の手法を用いて，文書間の類似度を求めるシステムについて提案した．

本提案では，膨大な文書を P2P におけるノードと考え，文書間類似度計算を削減するために，バネ系のシュミレーションとして知られている Vivaldi を用いた類似度予測を提案した．さらに，各文書を P2P 手法によってクラスタリングする手法を示した．

本システムを用いることで，計算量を大幅に削減しつつ，柔軟な類似度に基づく文書検索が可能になると考えられる．

Table2　クラスタ情報

クラスタの種類		保持する情報
所属クラスタ	末端以外	クラスタ ID，遠方ノードのリスト
	末端	クラスタ ID，所属する全ノードのリスト，遠方ノードのリスト
兄弟クラスタ		クラスタ ID, 代表ノード，所属ノード数

参考文献

1) Frank Dabek, Russ Cox, FransKaashoek, and Robert Morris.: Vivaldi: A Decentralized Network Coordinate System In the Proceedings of the ACM SIGCOMM '04 Conference, Portland, Oregon, August 2004
2) BambooDHT http://www.bamboo-dht.org/
3) 上田達也, 安倍広多, 石橋勇人, 松浦敏雄. P2P 手法によるインターネットノードの階層的クラスタリング, 情報処理学会論文誌, Vol.47 No.4
4) ISDL Report システム
http://mikilab.doshisha.ac.jp/dia/research/report/
5) Wikipedia
http://wikipedia.org/