様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　１０　日現在

機関番号：　１１３０１
研究種目：　基盤研究(C)
研究期間：　2012～2014
課題番号：　２４５００７２

研究課題名（和文）超高次元データの効率的な類似度検索を可能にする相補的Ｐ２Ｐ分散システムの開発

研究課題名（英文）Development of complementary P2P distributed system that enables an efficient similarity search of very high-dimensional data

研究代表者

菅谷　至寛（SUGAYA, YOSHIHIRO）

東北大学・工学（系）研究科（研究院）・助教

研究者番号：８０３２３０６２

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：Peer-to-peerシステムは膨大な情報を扱う上で有望な手段の一つだが，ネットワーク全体に対する類似度検索を行うことは困難な問題であり，既存手法は低次元データでのみ有効である．
本研究は高次元データに対して類似度検索を実現することを目的としている．非構造化オーバーレイネットワーク，および構造化オーバーレイネットワークそれぞれについて検討を行い，性能を実際のデータを用いた実験によって確認した．

研究成果の概要（英文）：Peer-to-peer system is a promising solution to manage huge amount of information, but similarity search over whole networks is a challenging problem. Existing methods that have ability of similarity search over whole network work only with low-dimensional data.
In this research, we aim at realizing similarity search for very high-dimensional data. We proposed methods using unstructured overlay network and structured overlay network, and confirmed the effectiveness by experiments using real data.

研究分野：　並列分散処理

キーワード：　オーバーレイネットワーク　類似度検索　Vivaldi

１．研究開始当初の背景

近年, ユーザー参加型の情報共有サービスが注目されている．非常に多くの利用者が情報を発信するようになり，コンテンツの総量は膨大となってきている．そのような情報の海から有益な情報を引き出す手段の一つとして，データセット全体から，あるデータに類似したデータを探すという「類似度検索」が挙げられる．例えば，「本棚.org」という，自分の蔵書セット（本棚）を登録して公開する実験的な情報共有サービスが存在するが，「類似した本棚を探す」ことが「類似度検索」に相当する．類似度検索は，関心がある，または関連すると思われる情報を探し出すために有用であり，他の情報共有サービスやSNS，ショッピングサイトなどでも，レコメンドの一種として同様の検索手法が利用されている．

大量のデータを低コストで保持するための手段の一つとして，P2P ネットワーク技術を用いることが考えられるが，類似度検索を含む柔軟な検索を P2P ネットワーク全体から行うことは容易ではない．様々な研究が行われているが，コンテンツが非常に高次元でスパースなベクトルで表される場合は特に困難で，未解決である．

２．研究の目的

本研究では, コンテンツが非常に高次元でスパースなベクトルで表される場合において，類似度検索を含む柔軟な検索をできるだけ少ないネットワーク負荷で実現するためのデータ表現方法，およびオーバーレイネットワークを開発することを目的とする．

３．研究の方法

類似度検索（Top-k 検索）では，あるコンテンツに対して類似度が上位のコンテンツだけが重要であり，明らかに似ていないコンテンツ間では類似度を計算する必要がない．したがって，類似コンテンツが近くのノードに配置されるようにデータ配置の局所性を高める工夫が重要となる．P2P ネットワークには主に非構造化オーバーレイネットワークと構造化オーバーレイネットワークがあるが，それぞれについて検討を行った．

(1) 非構造化オーバーレイネットワークによる方法

非構造化オーバーレイにおける探索はフラッディングが基本であり，類似度検索を含む柔軟な検索を行うことは，近傍に対しては比較的容易である．しかし，ネットワーク全体を探索するには多くのノードにクエリを転送する必要があるため，メッセージ量が多くなってしまう傾向がある．本研究では，検索精度を保ちながら通信量を抑えるために，検索クエリ転送の制御方法およびデータ配置の自己組織化の検討を行った．

(2) 構造化オーバーレイネットワークによる方法

構造化オーバーレイネットワークのうち, 一般的に広く用いられている分散ハッシュテーブルでは類似度検索を含む柔軟な検索が難しいものも多いが，ZNet や SkipIndex など, 完全一致検索だけではなく範囲検索が可能なものも存在する. 類似度検索は特徴空間での範囲検索とみなすことができるため, これを利用して類似度検索を実現することができる．

しかし, これらの手法はコンテンツが低次元な多値ベクトルで表される場合には有効だが, 非常に高次元のベクトルで表される場合や２値ベクトルで表される場合はそのまま適用することができない. 本研究では, コンテンツが非常に高次元でスパースなベクトルで表される場合を想定しており，例えば前述の「本棚.org」では，一つのコンテンツは一つの本棚, すなわち, 本のリストであり，（本１，…，本ｎ）のような長くスパースな２値ベクトルとして表現される．ここで, 各要素は当該リストでのそれぞれの本の有無を表し, ｎは全体での本の種類数となり非常に大きい．

そこで本研究では, 高次元でスパースな２値ベクトルによる空間を, コンテンツ間の距離を出来るだけ保ったまま，一旦, 低次元の空間に変換することを考える. その後, 低次元での範囲検索が可能な ZNet を用いることで類似度検索を実現する．

低次元空間への変換は, 全てのデータが一箇所に集められた状態であれば PCA (Principal Component Analysis) を始めとして様々な手法が研究されているが, 本研究ではデータが分散された状態であることを前提としている. したがって, 多くの通信を必要とする手法はスケーラビリティの観点から現実的でなく, 出来るだけ大域的な情報を必要としない方法でなければならない．本研究では，Counting Filter (CF) による方法と，Vivaldi による手法を検討した．

４．研究成果

(1) 非構造化オーバーレイネットワークによる方法

フラッディングによる検索では, 多くのピアがクエリを受信しなければ高い精度を維持することが困難である. そのため各ピアはより多くのピアにクエリを送信する必要があり, ネットワーク全体に大きな負荷がかかることとなる. この問題を解決するため, 過去に行われた検索結果を利用しクエリの送信先を制御することによって，トラフィックを軽減するとともに, 類似度の高いデータを持つピアへ確実にクエリを送信する手法を開発した．

提案手法では，（過去の）検索結果に基づき，各ピアが転送キャッシュと転送抑制キャッシュを作成･共有することでクエリ転送の制御を行う．転送キャッシュは包囲長方形とピアリストの組からなり，ピアリストに含まれるピアが包囲長方形内のデータを保有していることを記憶している．また，転送抑止キャッシュは包囲球とピアリストの組であり，ピアリストに含まれるピアが包囲球内のデータを持っていないことを表す．いずれのキャッシュも検索時にその結果に基づいてエントリが作成され，検索クエリの経路を逆にたどって経路上のノードで共有される．これらによって，過去の検索と類似する検索では，クエリを転送すべきノードを絞ることができる．さらに，検索データとその検索でヒットしたデータのうち最も類似度が低かったものを保持することで，より厳密な転送制御を行えるように改良を行った．

人工データ（ランダムデータ，クラスタリング可能なデータ）および本棚.org による実際のデータを用いてシミュレーション実験を行い，提案手法の性能を確認した．人工データに関しては，ランダムデータおよびクラスタリング可能なデータのいずれの場合でも，提案手法はフラッディングと同等程度の精度を得ながらメッセージ数を大幅に抑える（１／４程度）ことができた．しかし，本棚.org による実際のデータでは，提案手法によってメッセージ数は大幅に抑制できている（１／６程度）が，検索精度はフラッディングよりも低かった．図１，図２，図３に結果を示す．

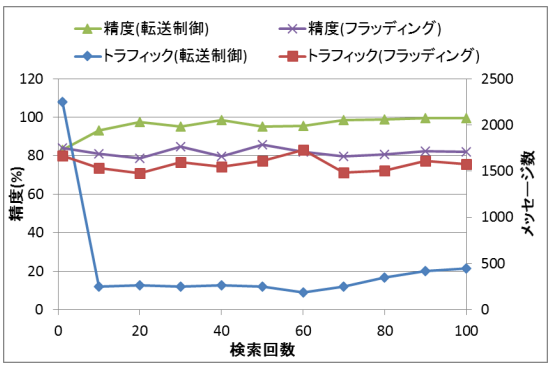

図 1　検索精度とメッセージ数（ランダムデータ）

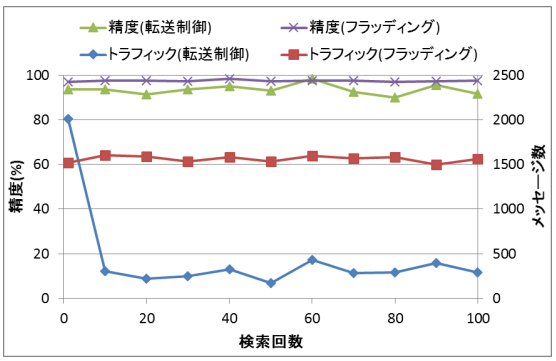

図 2　検索精度とメッセージ数(クラスタリング可能なデータ)

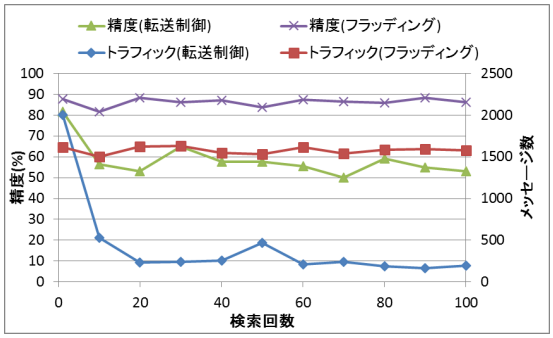

図 3　検索精度とメッセージ数（本棚.org）

(2) 構造化オーバーレイネットワークによる方法

データが分散化された状態であることを考慮し，大域的な情報をなるべく用いずに，コンテンツ間の距離関係をできるだけ保ったまま低次元の空間に射影するための方法として，CF による手法と Vivaldi による手法を検討した．

なお，データ間の距離関係をできるだけ保ったまま任意の次元の空間に埋め込む手法としては，多次元尺度構成法（MDS: Multi Dimensional Scaling）が良く知られている．本研究の初期段階では MDS を分散環境向けに適応させた手法も検討していたが，データによっては収束しない，または収束に非常に時間がかかるという問題が確認されたため，MDS の代わりに Vivaldi による手法の検討を行った．

① CF による手法

CF は Bloom Filter の２値カウンタを多値に拡張したもので，カウンタ値の列をベクトルとみなすことで，要素の集合を多値ベクトルとして表現することができる．したがって，これを用いることで，多次元でスパースな 2 値ベクトルを任意の次元の多値ベクトルに変換することができる．その際，元の空間で類似していたデータは，変換後の空間でも近い位置にあることが期待できる．ただしその性質上，類似していないデータが変換後の空間では近い位置になる場合もある．

この手法は LHS (Locality Sensitive Hashing) の一種と考えられ，変換時にはノード間の通信を全く必要としないという利点がある．

② Vivaldi による手法

Vivaldi はネットワーク座標系の一種で，本来は分散ネットワークにおいてノード間の距離を反映した仮想的な座標を推定するための手法である．ネットワーク全体をばねモデルとみなし，ノード間はばねによって接続されていると仮定する．ネットワーク全体のばねエネルギーを最小化することによって，極一部だけのノード間の距離を用いて任意の次元数における相対位置を推定

する．あるノードが他のノードと通信を行うとき，その 2 つのノードがばねでつながっているとみなして，推定座標から計算される距離と測定された距離（round-trip time）の差に基づき，逐次的に座標を更新していく．提案手法では，データをノードとみなすことによって高次元データの類似性に基づいた次元圧縮に応用した．

ばねで繋がれる相手ノード（データ）の数および選択方法は Vivaldi の収束速度や誤差に影響を与えると考えられるため，これについても検討を行った．その結果，現在の推定座標内での最近傍を相手ノードとして選択すれば良いことを実験によって確認した．

③ CF + Vivaldi

CF による手法と Vivaldi による手法は併せて用いることができる．CF によって得られた座標を初期値として Vivaldi を適用する．これによって，Vivaldi が収束するまでにかかる更新回数を大幅（１／２程度）に低減できることを実験によって確認した．

これらの手法（CF，Vivaldi，CF + Vivaldi）および比較手法として PCA を用いて 16 次元に次元圧縮を行った後にノード数 1024 の ZNet 上に配置し，検索範囲を変えた際の Top-10 の検索精度とメッセージ数の関係を確認した．類似度にはコサイン係数を用い，データには本棚.org による実際のデータを用いた．

結果を図４に示す．メッセージ数を少ない数に制限した場合は CF + Vivaldi の検索精度が最も高かった．メッセージ数と検索精度にはトレードオフがあり，メッセージ数を抑制することは分散環境では非常に重要となる場合がある．メッセージ数が多い場合では PCA の精度が最も高いが，PCA を計算するためには一般に全データを集約する必要があり，このことは分散環境では望ましくないことに注意する必要がある．また，メッセージ数が多い領域では CF だけでもある程度高い精度を達成できることを確認した．

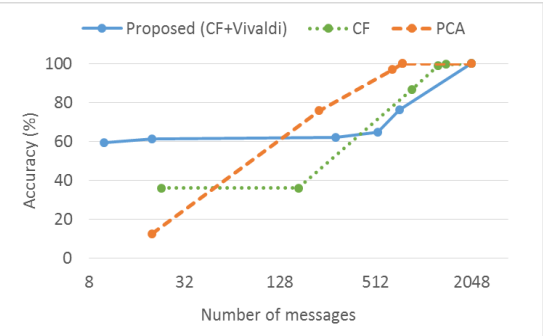

図 4　Top-10 検索の精度とメッセージ数

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計５件）

① Yoshihiro Sugaya, Koh Motoyama, and Shinichiro Omachi, Data Arrangement and Dimensional Compression using Vivaldi for Similarity Search on Structured Peer-to-peer Network, IEEE International Conference on Consumer Electronics - Taiwan, 2015 年 6 月 6 日，台北市（台湾）

② 本山洸，菅谷至寛，大町真一郎，構造化 P2P ネットワークにおける類似度検索のための次元圧縮手法の検討，マルチメディア通信と分散処理ワークショップ，2014 年 12 月 8 日，ホテル玉泉（島根県・松江市）

③ 本山洸，菅谷至寛，大町真一郎，構造化 P2P ネットワークを用いた高次元情報の探索，平成 26 年度電気関係学会東北支部大会，2014 年 8 月 22 日，山形大学工学部（山形県・米沢市）

④ 本山洸，菅谷至寛，大町真一郎，分散環境における類似度検索のための次元圧縮法の比較検討，電子情報通信学会 2014 年総合大会，2014 年 3 月 19 日，新潟大学（新潟県・新潟市）

⑤ Koh Motoyama, Yoshihiro Sugaya, and Shinichiro Omachi, A Study on Dimensionality Reduction for Similarity Search in Distributed Network, 2013 International Workshop on Emerging ICT, 2013 年 10 月 29 日，東北大学（宮城県・仙台市）

６．研究組織

(1)研究代表者
菅谷 至寛（SUGAYA, YOSHIHIRO）
東北大学・大学院工学研究科・助教
研究者番号：80323062

(2)研究分担者
大町 真一郎（OMACHI, SHINICHIRO）
東北大学・大学院工学研究科・教授
研究者番号：30250856

(3)連携研究者
なし