様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月 １４ 日現在

機関番号：１２６１２
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ１１３１１

研究課題名（和文）時間と共に変化する集合を対象とした類似検索

研究課題名（英文）Continuously Similarity Search for Evolvingi Sets

研究代表者
古賀　久志（Koga, HIsashi）

電気通信大学・大学院情報理工学研究科・准教授

研究者番号：４０３６１８３６

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：本研究では、ストリームデータを対象とした類似検索を取り扱った。ストリーム内の直近のデータを要素が入れ替わる集合としてモデル化し、集合に対する類似検索技術を拡張して、ストリームデータに対する類似検索を実現した。とくに類似検索結果を確定するのに必要な分だけデータ間で類似度計算を行い、不要な類似度計算を避けることで類似検索を高速化した。また、ストリームデータに対する類似検索問題を定式化したこと自体も本研究の成果である。

研究成果の学術的意義や社会的意義
本研究で定式化した類似検索問題は情報推薦という現実的な応用を持つ。例えばSNSにおいては、ユーザUのストリームをUが投稿したテキスト群でモデル化できる。この時、ストリームのスライディングウインドウはユーザUが直近に投稿したテキスト集合となり、Uが最近興味を持った事象を色濃く反映している。従って、スライディングウィンドウが似たユーザを探すことで、最近の興味が似たユーザを発見できる。そして、類似ユーザが閲覧しているニュース記事をお薦めするといった情報推薦サービスも実現可能になる。

研究成果の概要（英文）：This research studied similarity search for data streams. In particular, we regard the latest data in a data stream as an evolving set whose elements can change dynamically. Then, we realized the similarity search for data streams by reducing the problem to the set similarity search. In particular, we developed several fast similarity search algorithms that measure the similarity between two data just enough to determine the search results, avoiding unnecessary similarity computations. It is also our research contribution that we formulated two similarity search problems for data streams.

研究分野：データ構造、検索アルゴリズム

キーワード：類似検索　集合　ストリームデータ　スライディングウィンドウ　情報推薦

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

近年、IoT 等の普及により、ストリームデータに対する類似検索の重要性が高まっている。ここで、心電図や人間の動作の軌跡などのように「新たな計測値の追加により時間と共に変容するデータ」のことをストリームデータと呼ぶ。個々のストリームデータは数値の時系列、あるいは計測値がカテゴリ値である場合は文字系列として表現され、時間経過と共に数値や文字が追加される。

図１：ストリームデータ

ストリームデータを対象とする類似検索では、新しい情報を重視するため、系列データの直近の W 個の要素のみ(W は定数パラメータ)を用いてデータ表現するスライディングウィンドウモデル（図１）がよく用いられる。そして類似度は、ウィンドウ内の要素順を考慮できる時系列データ間類似度や文字列間類似度が使われて来た。

しかし、情報推薦への応用を対象とする場合、要素間の厳密な順序は重要ではない。例えば、ユーザ U に対して U の最近の嗜好に合った情報(ウェブニュースやウェブ広告等)を提示する情報推薦では、U が最近閲覧したニュースの厳密な時間順序は重要ではない。そこで、本研究ではスライディングウィンドウ内のコンテンツを時系列ではなく集合、とくに時間経過と共に一部の要素が入れ替わる集合として取り扱う。「時間と共に変化する集合」は情報推薦という現実的な応用を持つ重要なデータ構造であるにも関わらず、これまで研究対象とされて来なかった。従って、その性質を解明し、発見された性質を活用した効率的なアルゴリズムの実現が望まれる。

２．研究の目的

本研究では、時間と共に変化する集合に対する効率的な類似検索の実現を目的とする。とくに、データ追加に依る集合の変化に追従し、検索結果を随時更新する continuous な類似検索を高速処理するアルゴリズムを実現する。しかし現実には、既存研究が少ない研究テーマであるため、類似検索問題の妥当な定式化から始める必要がある。さらに２つの集合が似ているかどうかの判定も必要になるため、どのような集合間類似度を用いるかも指定する必要がある。

こうして類似検索問題が定義されると、問題に対する答えも定まる。continuous な類似検索を高速化するには、集合が変化する度に集合間類似度を再計算するのを止め、検索結果が変わる可能性がある時のみ集合間類似度を更新することが望まれる。従って、検索結果が変化する可能性の有無を見極める技術の確立も本研究の目的の一部となる。

我々が定式化した問題はいずれも、ユーザ U が新たにウェブニュースを読むことで嗜好が更新され、お薦めすべきニュースが変わる状況をモデル化しており、本研究は、ユーザ U の興味の変化に適応した情報推薦を実現する基盤技術となる。

３．研究の方法

本研究で、検索問題を定式化するに当たり、以下の２点を考慮する。つまり、これら 2 つを決定すると１つの検索問題が定まる。

(1) 時間と共に変化する（以下、動的に変化する）対象
(2) 集合要素のデータ型

(1)については、類似検索はクエリ(query)と似たデータをデータベースから探すというフレームワークであるため、(A)クエリのみが動的に変化、(B)データベース側のみが動的に変化、(C)クエリとデータベースの両者が動的に変化という 3 ケースが考えられる。(A)は従来研究で存在するため、(B)から(C)の順に研究対象を段々難しくすることで着実に研究を進める。

(2)については、集合の要素は通常ラベル（文字）である。しかし、SNS の場合、ユーザ U はテキスト（例：twitter でのツイート）を投稿することから、集合の要素がテキストとなるケー

スも取り扱う。
(2)をさらに発展させると集合の要素が木やグラフのように構造化されたデータであるケースも考えられる。そこで本研究では上記と並行してグラフ間類似度の改善にも取り組んだ。

4. 研究成果

(1) 動的に変化するデータベースを対象にした継続的な類似検索（CED 問題：Continuous similarity search for Evolving Database）

クエリ Q が静的集合で不変であり、データベースにn個のストリームデータ$\{S_1, S_2, \ldots, S_n\}$が登録された状況で、Q と最も類似した上位 K 個の集合を求める検索問題を取り扱った。ストリームデータに新データが到着すると集合の要素が変化するので、検索結果を continuous に更新する必要がある。集合 Q とS_i（のウィンドウ）間の類似度は Jaccard 係数 （式(1)）を採用した。

$$\frac{|Q \cap S_i|}{|Q \cup S_i|} \tag{1}$$

以下ではこの問題を CED 問題と呼ぶ。CED 問題のアプリケーションは、ウェブ広告をどのユーザに提示するかという情報推薦である。ウェブ広告をそのカテゴリを示すキーワード集合で表し、各ユーザ U をウェブページの閲覧履歴ストリームで特徴表現する状況を考える。この時，ストリームのスライディングウィンドウ内の履歴は U の直近の嗜好を表す。その上で、ウェブページもカテゴリを表すキーワード集合と紐付ければ、ウィンドウ内の履歴は時間と共に変化するキーワード集合となる。そして，ウェブ広告をクエリ Q とすることにより、ウェブ広告 Q と嗜好性が適合する上位 K 人のユーザを発見できる。

CED 問題に対して、我々は将来の類似度を予測することで不要な類似度計算を抑制する高速アルゴリズム CE(Common Element)を考案した。具体的には現時刻の類似度が式(1)である時、時刻 T 後には集合の要素がせいぜいT個しか変化しないことから、類似度が式(2)以下であることが言える。

$$\frac{|Q \cap S_i| + T}{|Q \cup S_i| - T} \tag{2}$$

さらに、集合$Q \cap S_i$内の共通要素はS_iのウィンドウ内に現存することから、時刻Tまでに共通要素が何個ウィンドウから離脱するかも把握できる。その数を v とすると、時刻T後の類似度の上限値をより厳密に求められる(式(3))。

$$\frac{|Q \cap S_i| + (T - v)}{|Q \cup S_i| - (T - v)} \tag{3}$$

式(3)の上限値が小さいストリームは類似度が上位 K 位に入る可能性がないので、類似度計算を省略できるというのが CE 法のアイデアである。CE 法で上限値が増加する様子を図 2 に示す。横軸が経過時間 T で縦軸が類似度上限値である。CE 法では共通要素が離脱する時刻に上限値が増えず、上限値の増加スピードが遅くなるため、類似度を再計算するまでの期間を伸ばせている。

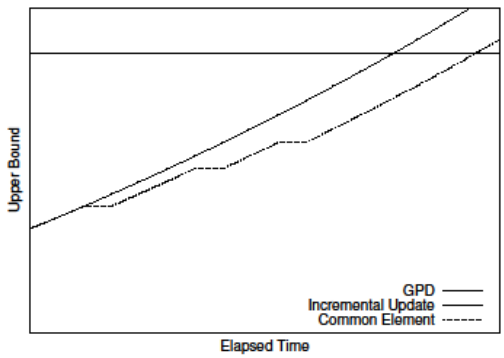

図 2: 類似度上限値の経過(式(3):Common Element，式(2): Incremental Update)

実データを用いた実験により CE 法が式(3)とは別の上限値を利用したベースライン手法より実行時間を約 30%削減できることを確認した。

(2) 動的に変化するテキスト集合に対する類似検索(CTS 問題，Continuous similarity search for Text Sets)

ここでは twitter のような SNS で、ユーザ U を SNS に投稿したテキスト群でモデル化し、類似テキスト集合を探すことにより U に対する類似ユーザを探す問題を考えた。この問題を CTS 問題と呼ぶ。CTS 問題では、クエリユーザ Q のスライディングウィンドウをクエリテキスト集合とする。クエリテキスト集合は Q が最近投稿したテキストの集合であり、Q の最近の興味を反映している。同様に、データベースには Q とは異なる n 人のユーザのテキストストリームが登録されていて、それらのスライディングウィンドウは各ユーザの最近の興味を反映する。従って、Q とテキスト集合が似たユーザを探せば、最近の興味が似たユーザを発見できる。

しかし 2 つのテキスト集合が似ているかどうかを数値化するのは簡単でなく、テキスト集合に対して類似度を定義する必要がある。時刻 T でのユーザ Q とユーザUのスライディングウィンドウQ_T，U_TはどちらもW個のテキストで構成される。Wはウィンドウサイズである。我々は、Q_T，U_Tの各テキストを頂点とし、共通単語を持つテキストペア間に枝を張った 2 部グラフ$G(Q_T, U_T)$を考え、その極大マッチングサイズを類似度と定めた。CTS 問題は、Q との類似度が閾値 ϵ 以上となる全ユーザをデータベースから列挙する継続的なレンジ探索である。そして、スライディングウィンドウが前進する度に検索結果を更新することが要求される。

我々は CTS 問題に対して枝刈りベースの高速アルゴリズム（遅延評価法）を考案した。CTS 問題での類似度は極大マッチングサイズであるが、遅延評価法では極大マッチングを完成させず、ユーザUの類似/非類似を判定するのに必要な分だけ極大マッチングを部分的に完成させる。極大マッチングを完成させずテキスト比較回数を削減することで、遅延評価法は判定処理にかかる時間を短縮する。より具体的には遅延評価法ではQ_T内のW個のテキストを

① Q_M: 極大マッチングに含まれることが確定したテキストの集合
② Q_{NM}：極大マッチングには含まれないことが確定したテキストの集合
③ Q_{UM}：極大マッチングに含まれるかが未定のテキストの集合

という 3 グループに分けて管理する。時刻が T-1 から T に進んだ時、遅延評価法は Q，U のスライディングウィンドウに入ってくるテキスト INv，INu，スライディングウィンドウから離脱するテキスを OUTv，OUTu を処理した後、$|Q_M| > \epsilon$となって類似と判定される、あるいは$|Q_{NM}| > W - \epsilon$となって非類似度と判定されるまで、Q_{UM}に属するテキストのマッチング判定を繰り返す（図 3）。

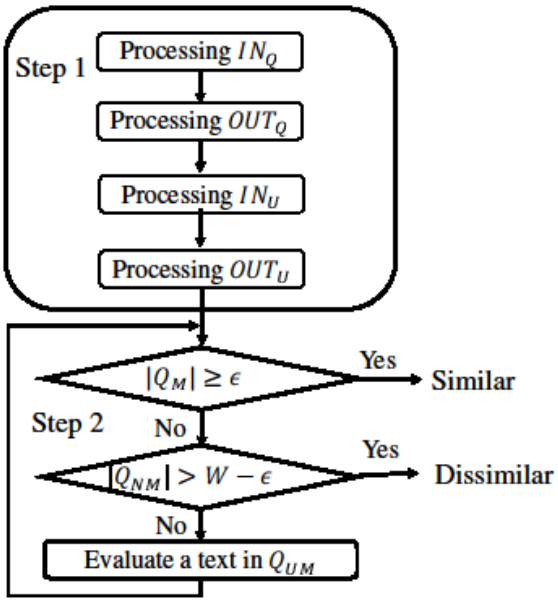

図 3：遅延評価法の処理

実データを用いた実験により遅延評価法が極大マッチングを完成させるベースライン手法より実行時間を短縮できることを確認した。短縮幅は類似度閾値 ε に依存する。

(3) ライングラフを用いたグラフのベクトル表現

グラフはノードとエッジ（辺）で構成されるデータ構造であるが、2 つのグラフが似ているか

どうかを数値化するのは簡単ではない。つまり、グラフ間類似度を算出するのは自明ではない。この状況に対して、ニューラルネットワークの発展に伴い、グラフをベクトル化してベクトル間で距離計算してグラフ間距離を求める方式が出現している。有名な手法としては Graph2vec がある。Graph2vec はノードにラベル（属性）が付与されたグラフをベクトル化できる。しかし、エッジラベルは取り扱うことができない。エッジラベルは、例えば化合物の場合、水素結合や共有結合といった結合タイプを表し、分類に有用な情報を持つ。

我々は Graph2vec を、エッジラベルを取り扱えるように拡張した。具体的にはグラフ G に対するラインググラフ L(G)を利用した。ラインググラフ L(G)は元のグラフ G に対する双対グラフであり、G の辺が頂点となる（図４）。G のエッジラベルが L(G)の頂点ラベルとなるため、L(G)を Graph2vec でベクトル化すれば、G のエッジラベル情報を含んだベクトル表現を生成できる。

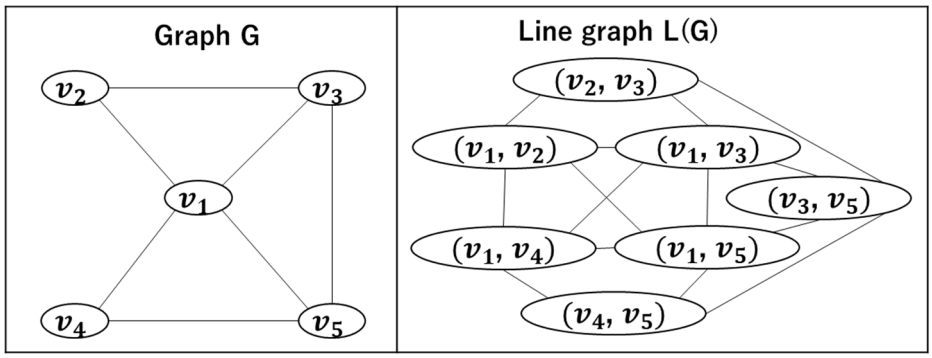

図 4：グラフGとそのラインググラフL(G)

最終的には G のベクトル表現と L(G)のベクトル表現を連結することで、G のノードラベル情報とエッジラベル情報の両者を含んだベクトル表現を得る。提案手法を GL2vec(Graph and Line graph to vector)と名付けた。複数個のデータセットを用いてクラス分類実験を行ったところ、GL2vec の分類精度が Graph2vec の分類精度を上回った（表１）。この事実より GL2vec の有効性を確認した。

表 1. 分類精度 (mean ± std dev.)

datasets	MUTAG*	NCI33	NCI83	DBLP
Graph2vec	83.68 ±7.02	78.95 ±1.82	75.90 ±1.66	90.63 ±0.59
GL2vec	**87.63** ±7.50	**81.30** ±2.17	**77.29** ±1.31	**92.27** ±0.62

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 H. Koga and D. Noguchi	４．巻 12440
２．論文標題 Continuous Similarity Search for Evolving Database	５．発行年 2020年
３．雑誌名 springer LNCS, Proc. 13th International Conference on Similarity Search and Applications(SISAP 2020)	６．最初と最後の頁 155-167
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-60936-8_12	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 陳 宏, 古賀久志	４．巻 62(1)
２．論文標題 非教示なグラフ分散表現のエッジ特徴による改良	５．発行年 2021年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 357-368
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 藤原 勇二, 古賀久志	４．巻 62(3)
２．論文標題 多観点類似度を用いた凝集型階層クラスタリング	５．発行年 2021年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 936-945
掲載論文のDOI（デジタルオブジェクト識別子） 10.20729/00210263	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Y. Fujiwara and H. Koga	４．巻 11707
２．論文標題 Multiviewpoint-Based Agglomerative Hierarchical Clustering	５．発行年 2019年
３．雑誌名 springer LNCS, Proc. 19th International Conference on Database and Expert Systems Applications(DEXA 2019)	６．最初と最後の頁 325-340
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-27618-8_24	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 H. Chen and H. Koga	4．巻 11955
2．論文標題 GL2vec: Graph Embedding Enriched by Line Graphs with Edge Features	5．発行年 2019年
3．雑誌名 springer LNCS, Proc. 26th International Conference on Neural Information Processing(ICONIP 2019)	6．最初と最後の頁 3-14
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-36718-3_1	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 Hisashi Koga, Satoshi Suzuki, Taiki Itabashi, Gibran Fuentes Pineda, Takahisa Toda	4．巻 11314
2．論文標題 Extended Min-Hash Focusing on Intersection Cardinality	5．発行年 2018年
3．雑誌名 Springer LNCS, Proc. 19th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2018)	6．最初と最後の頁 17～26
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-03493-1_3	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 久保幸平，古賀久志
2．発表標題 ストリーム環境でのテキスト集合に対する類似検索
3．学会等名 情報処理学会数理モデル化と問題解決(MPS)研究会
4．発表年 2020年

1．発表者名 久保幸平，古賀久志
2．発表標題 ストリーム環境での位置情報を持つテキスト集合に対する類似検索
3．学会等名 第18回情報科学技術フォーラム(FIT2019)
4．発表年 2019年

１．発表者名
野口大樹，古賀久志

２．発表標題
動的に進化するデータベースを対象にした継続的な類似検索

３．学会等名
第12回データ工学と情報マネジメントに関するフォーラム(DEIM2020)

４．発表年
2020年

１．発表者名
野口大樹，古賀久志，戸田貴久

２．発表標題
集合間類似度を用いたストリームデータのtop-k類似検索における枝刈アルゴリズムの改善

３．学会等名
第17回情報科学技術フォーラム(FIT2018)

４．発表年
2018年

１．発表者名
土田祐将，古賀久志

２．発表標題
転置インデクスを用いた動的なテキスト集合に対する類似検索の高速化

３．学会等名
電子情報通信学会データ工学(DE)研究会

４．発表年
2021年

１．発表者名
土田祐将，古賀久志

２．発表標題
動的なテキスト集合に対する類似検索アルゴリズムALE-Qの評価

３．学会等名
第14回データ工学と情報マネジメントに関するフォーラム(DEIM2022)

４．発表年
2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関