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 ２８ 年　 ６ 月　 ９ 日現在

機関番号：１２６１２
研究種目：基盤研究(C)（一般）
研究期間：2012～2015
課題番号：２４５００１１１

研究課題名（和文）構造を持つデータの構造情報ダイジェストを利用した類似検索の高速化

研究課題名（英文）Fast Similarity Search for Structural Data using Structural Digests

研究代表者
古賀　久志（Koga, Hisashi）

電気通信大学・情報システム学研究科・准教授

研究者番号：４０３６１８３６

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：本研究では、グラフのような構造情報を持つデータを対象とする類似検索問題に取り組んだ。構造を持つデータは複雑度が高く、類似度計算のオーバーヘッドが問題になる。そこで、構造情報を要約する簡潔なデータ構造（ダイジェスト）を導入し、ダイジェストを比較することで、類似検索を高速化する。ここで、検索精度は構造情報をどう要約するかに依存する。本研究では、部分構造同士が似ているという情報をダイジェストに反映させることで、高精度な類似検索を実現した。
また、構造を持つデータに対する類似検索の応用技術として、構造ベースの画像処理についても研究した。

研究成果の概要（英文）：This project studied the similarity search for structural data such as graphs. Because structural data are complex, the overhead to compute the similarity between two data tends to be enormous. Therefore, we generate a small data structure called "digest" for an individual graph and measure the similarity between two data in a short time by comparing their digests. Here, the search accuracy depends on how to summarize the structural information of a graph onto its digest. By embedding the similarity information regarding substructures into the digests, we succeeded in realizing the similarity search with a high retrieval accuracy.

We also studied the structure-based image processing which is a promising application of similarity search for structural data.

研究分野：類似検索

キーワード：グラフ間類似度　類似検索　ハッシュ　構造ベース画像処理

１．研究開始当初の背景

類似検索はパターン認識において必須となる基盤技術である。一般に、類似検索では類似度計算の計算量が大きい。したがって、高速な類似検索を実現するには、軽量な手法で類似している可能性がないデータをあらかじめ多く排除し、類似している可能性があるデータ間でのみ類似度計算を行うことが重要になる。

とくに本研究課題では、木やグラフのような構造を持つデータ（以下、構造データと呼称する）に対する類似検索を対象とする。構造データは構造を持つことで高い表現能力を持つ。例えば、1 枚の画像に含まれる複数の特徴点は、特徴点を頂点とする属性付き平面グラフとして自然に記述できる。これは画像を単に特徴点の集合として表現するのに比べると、枝により特徴点間の近さのような空間情報を持たせることができるため記述力が高い。その反面で、構造データは類似度定義が単純ではなく類似度計算のオーバーヘッドが大きいという欠点も持つ。従って、構造データに対する類似検索では、類似している可能性がないデータをより効率的に除去することが求められる。よく見られる手法では、2 つの構造データの類似性をノードラベルの比較により判定して絞り込む。しかし、これでは構造情報を使っていないので絞り込み精度が悪く、とくに頂点に付与されるラベルの種類数が少ない場合に絞り込みができない。

２．研究の目的

本研究課題では構造データに対し、ラベル情報に加えて構造情報も畳み込んだサイズの小さい要約（以降、ダイジェストと呼ぶ）を生成し、ダイジェストが似ていないデータを類似可能性がないデータとして除去する手法を実現する。本手法により、ダイジェストに含まれる構造情報により、ノードラベルが似ていても構造の違いから類似しないデータ間での類似度計算を回避できることになる。ダイジェストの生成方法が類似検索性能に影響するので、構造情報をどうダイジェストに反映させるべきかを明らかにすることを研究目的とする。

上記と並行して、構造ベースパターン認識に関する研究も実施する。具体的には画像をグラフ表現し、パターン認識をグラフの類似性判定に帰着して実現することを目指す。

３．研究の方法

対象とする構造データを頂点にラベルが付与されたラベル付きグラフに絞り、ダイジェスト生成方法を検討する。提案手法の評価は、化合物データベースを用いて実施し、類似検索の精度を測定する。

構造ベースパターン認識の研究に関しては、主にグラフベース画像認識を取り扱う。画像を、特徴点/画素を頂点、空間的に近い特徴点/画素のペアを辺とするグラフとして表現して、パターン認識を実現する。こちらについては複数のアプリケーションを取り扱う。

４．研究成果

(1) ラベル付きグラフに対するダイジェスト生成：

提案手法におけるダイジェスト生成手法の流れは次のようになる。まず、ラベル付きグラフ G を、G の部分構造の集合として特徴表現する。具体的には、各頂点を始点とする有限長パスを部分構造とした。すなわち、G を内包するパスの集合として表現する。この時点で、グラフ類似検索はパス集合間の類似検索に帰着されるが、パス集合間の類似検索も自明ではない。そこで、パスを整数にマッピングする関数を用意し、グラフを整数集合として表現する。整数集合に対する類似検索に関しては、min hash fingerprint と呼ばれる効率的なダイジェスト生成手法が知られており、それを利用して G のダイジェストを生成する。min hash では整数集合からサンプリングしてダイジェストを生成するので、結局、G 内のパス集合からサンプリングして整数に変換したものがダイジェストになる。

2 個のグラフがどれだけ似ているかはダイジェストを比較することで見積もれる。ダイジェストを用いて、グラフの類似度を判定する処理の流れを図 1 に示す。

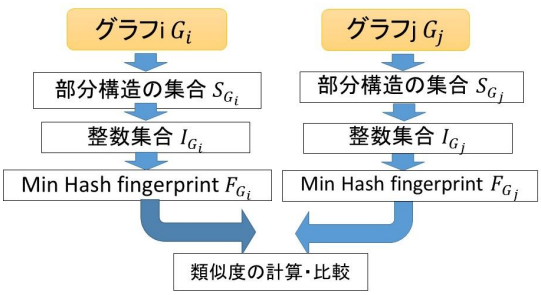

図 1：ダイジェストを使ったグラフの比較

本研究ではとくに、パスを整数にマッピングする関数の設計を工夫した。従来手法ではパスを疑似乱数を用いて整数にマッピングしていた。このため、ダイジェストを比較しても、2 つのグラフがどれだけ同一パスを含

むのかという情報が得られるだけであった。提案手法では、類似パスを同じ整数にマッピングすることで、グラフペアがどれだけ類似パスを含むのかをダイジェストから判別できるようにした。具体的には長さ n のパス (v1,v2,...,vn)を以下の式を用いて整数に変換する。

$$a1*v1+a2*v2+...+0*vc+.....+an*vn. \quad (1)$$

ここで ai は確率的に定めたパスの i 番目の頂点 vi に対する係数である。とくに vc の係数 ac=0 であり、これによりラベルが１つだけ異なる類似パスが同じ整数にマッピングされる。

化合物データベース PTC を用いた類似検索実験により提案手法を評価した。PTC は 417 個の化学化合物で構成されるデータベースである。化合物は元素を頂点、元素間の結合を辺とするグラフとして表現されている。類似検索実験の手順は以下のようになる。

1. PTC の要素であるグラフ G に対して、頂点を削除/置換して改変したグラフ G'を生成する。
2. G'をクエリとして最もダイジェストが似たデータベースを PTC から検索する。
3. 検索結果がオリジナルグラフ G と一致した時に類似検索が成功したと判定する。

G を変えて類似検索を繰り返し、正解率により、ダイジェストの性能を評価した。改変する頂点の割合を変えた時の従来手法と提案手法の正解率を表１に示す。

表 1 ：類似検索の正解率

	変更率	10%	20%	30%	40%
ラベル置換	提案法	0.92	0.66	0.44	0.32
	従来法	0.89	0.55	0.36	0.24
頂点削除	提案法	0.93	0.69	0.47	0.26
	従来法	0.93	0.68	0.45	0.26

表１より提案手法では、ラベル置換による改変を加えた時に正解率が従来手法を有意に上回った。一方で、頂点削除に関しては従来手法と提案手法の正解率はあまり変わらなかった。

以上より、提案手法では，ラベル置換操作に対してロバスト性の高いダイジェストを生成することに成功した。これは式(1)により、ダイジェスト生成時にラベルが僅かに異なるパスを同一整数にマッピングした効果である。

（２）グラフベースの画像オブジェクト発見

近年，デジタルカメラの普及により，扱う画像データの量が膨大になっている。このような背景から大量の画像を与えられた時に，人手によらない意味付け(アノテーション)機構へのニーズが高まっている。画像から特定オブジェクトを見つけるオブジェクト認識は盛んに研究されているが、探索対象となるオブジェクトモデルを人間が提示する必要がある。これでは画像の多様性が増加してくると対応が困難になる。そこで、多様なオブジェクトを含んだ画像群から、モデルを与えずに自動的にオブジェクトを発見する技術が重要になる。

本研究では、大量画像からオブジェクトモデルを自動発見する手法を構築した。提案手法では、１枚の画像を SIFT 特徴点を頂点とするドロネー三角形分割によりグラフ化し、グラフ集合(=画像集合)に対してグラフマイニング技術を活用してオブジェクトモデルを発見する。このようなアプローチ自体は既存研究にも見られるが、既存研究では画像内にオブジェクトが１つしか存在しないことを仮定し、単にグラフをクラスタリングするだけでオブジェクトモデルが発見している。本研究では、複数のオブジェクトを含む画像からオブジェクトモデルを発見できる点が新しい。

提案手法では、n 枚の画像をグラフ化して得られる G1,G2,...,Gn までの n 個のグラフを入力とする。グラフの各頂点には画像特徴を表すラベルが付与されている。本手法では、以下の原理に基づいて、オブジェクト自動発見を実現する。

- 同一オブジェクトが存在する位置には同じグラフ構造が出現する。
- 従って、同一オブジェクトに含まれる辺同士はそのオブジェクトが出現するたびに共起する。
- 逆に、何度も繰り返し共起する辺を凝集すれば、同一オブジェクトに所属する可能性が高い。

そこで、提案手法では，「繰り返し共起する辺の集合」を発見して、オブジェクトモデルとする。提案手法は、繰り返し出現するパターンを偶然の産物でない意味のある実体として取り出すので、頻出パターンマイニングの１種と位置づけられる。繰り返し共起する辺の集合を以下の 3 ステップで抽出する。

[Step1] 頻出辺の発見：ここでは、繰り返し共起する辺の必要条件として，繰り返し出現するつまり頻出辺を探索する。ここで、「辺が繰り返し出現する」とは両端点のラベルが同じ辺が複数個出現するということである。

[Step2] 共起する頻出辺ペアの発見：ここでは、頻出辺間で共起性を調べて共起する頻出辺ペアを発見する。頻出辺 ej に対して、それが出現した画像の集合を Sj とする。2 個の

頻出辺 ej と ek の共起性は、出現画像集合の Jaccard 係数(式(2))により測れる.

$$|Sj \cap Sk| / |Sj \cup Sk|. \quad (2)$$

この値が閾値θ以上になった時、ej と ek は共起性が高いと判定する。

[Step3] 共起する頻出辺ペアのクラスタリング：ここでは、共起辺ペアをクラスタリングすることにより、繰り返し共起する辺の集合をオブジェクトモデルとして取り出す。このクラスタリングは、共起辺がノードで、共起性があるノード間に辺が張られたグラフを、連結成分分解することで実現する。1 連結成分が 1 クラスタに対応する。連結成分の抽出はグラフの深さ優先探索により容易に実現できる。

提案手法の有効性を確認するための実験を行った。使用する画像は多視点画像データベース Columbia Object Image Library(COIL)を用いた。COIL は 100 種類のオブジェクトが収録されており，各オブジェクトに対して 5 度ずつ回転させた 72 枚の画像を含む合計 7200 枚の画像データベースである。

実験では COIL データベースに含まれる 2 枚の画像を連結させて複数オブジェクトを含んだ画像を生成した。画像例を図 2 に示す。画像内で、左には頻出オブジェクト、右にはランダムに選択したレアオブジェクトを配置した。左のオブジェクトを発見することが目的である。頻出オブジェクトの頻度は 1 種類あたり 10 とし、頻出オブジェクトの種類数を 30 とした。従って、実験画像データベースの画像枚数は 30×10=300 枚である。頻出オブジェクトの各インスタンスは角度が異なる。

図 2：実験に使用した画像

実験の結果、32 種類のオブジェクトモデルが発見された。これは頻出オブジェクトの種類数を上回るが、1 つの頻出オブジェクトが複数のオブジェクトモデルに分離されて発見されたためである。本手法では、教示を一切しないため、この現象はある程度許容されると我々は考えている。

発見されたオブジェクトモデルを使って、300 枚の画像に対するオブジェクト認識実験を行った結果、適合率は 1、再現率は 0.97 と良好な結果が得られた。これは提案手法が適切なオブジェクトモデルを獲得したことを示している。オブジェクト認識結果の一部を図 3 に示す。この図では、異なるオブジェクトモデルを異なる色で表しており、適切にオブジェクトが認識されたことが確認できる。

図 3：オブジェクト認識結果

（3）構造情報に着目した類似画像検索

本研究では類似画像検索を取り扱う。この分野では、Bag of Visual Words (BoVW) を用いた手法が数多く提案されている。BoVW では、画像内に存在する局所特徴量の分布をヒストグラムとして表現し、ヒストグラム類似度により画像の類似度を計算する。通常、BoVW では、画像内の特徴点をすべて均等に取り扱う。

一般に、画像は前景（オブジェクト）と背景という構造を有するが、類似画像検索においては、画像の意味（セマンティクス）は前景によって表されることが圧倒的に多い。そこで画像を前景と背景に分離して、前景ヒストグラムと背景ヒストグラムを作り、前景類似度と背景類似度を別々に求めることで、類似検索の性能向上を試みるアプローチがいくつか提案されている。

このような手法では、前景位置の推定が必要だが，色，コントラスト，エッジなどの特徴から顕著特徴マップを計算し、顕著度が高い特徴点は前景に所属し、顕著度が低い特徴点は背景に所属すると考えるのが一般的である。

しかしながら、顕著特徴マップは前景と完全には一致せず、誤りを含む。つまり、顕著度が低い特徴点が前景に所属することはそれなりの頻度で発生する。そこで、本研究ではこの事実に着目した類似検索方式を提案した。提案手法の独創的な点は以下の 2 つである。

1. 特徴点間の空間的な位置関係に着目し、顕著度の低い特徴点であっても、周辺の特徴点の顕著度が高い場合は前景に所属すると判定する。

2. 従来手法では顕著特徴マップから前景ヒストグラムと背景ヒストグラムを計算していた。提案手法では、前景ヒストグラムを計算するが、背景ヒストグラムは計算しない。その代わりに画像全体の特徴点から大域ヒストグラムを計算する。

背景ヒストグラムは顕著特徴マップの精度に依存するという理由で、提案手法では背景ヒストグラムを計算しない。従来手法では顕著特徴マップの精度が悪いと、前景ヒストグラムと背景ヒストグラムの両方が不正確になる。提案手法では、大域ヒストグラムが顕著特徴マップと無関係なので、顕著特徴マップの精度に関してよりロバストである。

Caltech-101 画像データベースを使用した評価実験を行った。Caltech-101 の 40 種類のクラスに対して、25 枚ずつ画像をランダムで選択し、計 1000 枚の実験用データセットを作成した。leave-one-out 法で類似検索実験を行った。類似検索結果に kNN 分類器を適用して、クエリ画像をクラス分類し、クエリ画像が正しいクラスに分類された時に類似検索が正解したと判定する。類似検索手法の性能は正解率によって評価する。

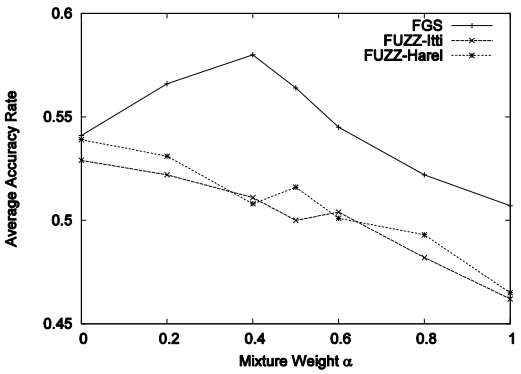

図 4： 類似検索の正解率

実験結果を図 4 に示す。このグラフでは FGS が提案手法で、FUZZ が従来手法である。縦軸は正解率、横軸が 2 種類のヒストグラム間類似度の混合率を表す。例えば、従来手法では、前景類似度と背景類似度の混合率であり、提案手法では、前景類似度と大域類似度の混合率である。グラフから提案手法が従来手法の性能を上回ることがわかる。

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）

〔雑誌論文〕（計　３件）

① Z. Zou and H. Koga, "Spatially Aware Enhancement of BoVW-Based Image Retrieval Exploiting a Saliency Map", in Proc. Computer Analysis of Images and Patterns (CAIP'15), springer LNCS Vol. 9257, pp.73-84, 2015. 査読有

② T. Nanbu and H. Koga, "Graph-Based Object Class Discovery from Images with Multiple Objects", in Proc. 15th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'14), springer LNCS Vol. 8669, pp.344-353, 2014. 査読有

③ S. Kudo, H. Koga, T. Yokoyama and T. Watanabe, "Robust Automatic Video Object Segmentation with Graphcut Assisted by SURF Features,", in Proc. 19th IEEE International Conference on Image Processing (ICIP 2012), IEEE, pp.297-300, 2012. 査読有

〔学会発表〕（計　６件）

① 宮田昂充, 古賀久志, 戸田貴久,"部分構造の類似性を考慮した Min-Hash ベースのグラフ類似検索",電子情報通信学会総合大会 D-4-11, 2016. 九州大学（福岡県福岡市）

② 土田和生, 古賀久志, 戸田貴久, "2 段階グラフカットを用いた動画からの移動物体抽出" 情処研報 2015-CVIM-196(20) 2015. 東北大学（宮城県仙台市）

③ 鄒子君, 古賀久志,"顕著特徴領域を利用した BoVW ベース類似画像検索の改善方式の検討", 信学技報 PRMU2013-110, pp.183-188, 2014. 大阪大学（大阪府豊中市）

④ Zhichao Zhang, Hisashi Koga, Youhei Ogyu, "Discovering NBA Game Stories from Twitter", 第 6 回データ工学と情報マネジメントに関するフォーラム (DEIM2014) 2014. ウェスティン淡路（兵庫県淡路市）

⑤ 南部貴之, 古賀久志, 渡辺俊典, "グラフの共起性に着目した複数オブジェクトを含む画像からの自動オブジェクト発見", 信学技報 PRMU2012-209, pp.169-174, 2013. 電気通信大学（東京都調布市）

⑥ 千田 祐真, 古賀 久志, 渡辺 俊典, "ビデオストリームからの対象ビデオの短時間検出", 第 5 回データ工学と情報マネジメントに関するフォーラム(DEIM2013) 2013. ホテル華の湯（福島県郡山市）

６．研究組織

(1)研究代表者

古賀　久志（Koga Hisashi）
電気通信大学・大学院情報システム学研究科・准教授
研究者番号：４０３６１８３６