様式 C-19

科学研究費補助金研究成果報告書

平成２４年５月２５日現在

機関番号：10101
研究種目：若手研究（B）
研究期間：2009～2010
課題番号：21700106
研究課題名（和文）時間変化するオブジェクト情報のＷｅｂからの収集と管理方式の研究

研究課題名（英文）Collection and Management of Information on Time-varying Objects on the Web

研究代表者
小山 聡（OYAMA SATOSHI）
北海道大学・大学院情報科学研究科・准教授
研究者番号：30346100

研究成果の概要（和文）：Web から人物や組織などのオブジェクト単位で情報を検索できるオブジェクトレベル検索エンジン実現のための重要な技術的課題の一つとして，オブジェクトの時間変化への対応がある．本研究では，時間変化するオブジェクトの同一性判定問題を時間的に離れたデータ間の関係を予測する時間横断的リンク予測問題として定式化し，機械学習を用いて精度の良い予測を実現する方法の開発などを行った．

研究成果の概要（英文）：Dealing with temporal variations of objects such as people and organizations is one of the important technical challenges to enable object-level search engines that treat an object as a retrieval unit. The contributions of this research include formalization of identifying time-varying objects as a cross temporal link prediction problem that infer the links among data in different time periods and development of a machine learning method that can enable accurate predictions.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,800,000	540,000	2,340,000
２０１０年度	1,500,000	450,000	1,950,000
２０１１年度	0	0	0
２０１２年度	0	0	0
２０１３年度	0	0	0
総　計	3,300,000	990,000	4,290,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：Ｗｅｂ オブジェクト検索 オブジェクト識別 時間情報 機械学習 リンク予測 リンク解析 情報抽出

１．研究開始当初の背景

人物や企業などの実世界のオブジェクトに関する情報は，Web 検索エンジンにおいて最も頻繁に検索される対象の一つである．しかし，たとえば人名を汎用の検索エンジンにクエリとして入力したとしても，目的の人物の情報が複数のページに分散していたり，同姓同名の人物の情報が検索結果に混在していたりといった問題が存在し，利用者が必要な情報を検索結果から見つけ出す際に負担となる場合が多い．そこで近年，複数の Web ページから，オブジェクト単位で属性や関係

（例えば人物であればメールアドレスや所属）を抽出してデータベースに格納し，その結果をユーザに呈示するオブジェクトレベル検索エンジンの研究が進められている．この中には，Microsoft の学術情報検索エンジン Libra(http://libra.msra.cn/)のように，サービスとして運用されているものもある．これらの研究の基本的なアプローチは，(1)ドメインのスキーマによりオブジェクトの持つ属性や関係を予め定義する(2)ヒューリスティックなパターンや機械学習によって得られた情報抽出規則でページから属性や関係を抽出する(3)同一オブジェクトに関する情報をクラスタリングによって識別し（オブジェクト識別やレコード同定等と呼ばれる）重複を除いて集約するというものである．我々もこれまでに，主に人物を対象オブジェクトとし，Web からの情報抽出やオブジェクト識別に関する研究を行ってきた.これらの従来研究を踏まえ，Web からのオブジェクト検索における重要な研究課題の一つとして，我々はオブジェクトの時間変化への対応があると考える．

２．研究の目的

人物などの実世界のオブジェクトは，その属性や他のオブジェクトとの関係が，時間とともに変化する．情報が上書きで更新される管理されたデータベースと異なり，Web においては過去の情報が残ったまま，新しい情報が追加されていく場合が多い．例えば，ニュース記事に現れる会社の代表者の名前や所在地などは，そのニュース記事が作成された時点での情報であり，現在の情報と一致するとは限らない．このように，オブジェクトの属性値や関係が時間変化することにより，古い情報を誤って提示したり，同一オブジェクトを別オブジェクト誤って判定したりといった問題が生じる．

本研究課題の目的はこのような問題を解決し，時間変化するオブジェクト情報の Web からの効率的な収集と管理を可能とすることである．

３．研究の方法

我々はこれまでの研究で，人物や企業の属性値の時間変化の確率（例えば，ある人物が一定時間に所属を変える確率）を考慮することで，オブジェクト識別精度を改善できることを示した．しかしこの方法では問題領域の知識を持つ設計者が人手で属性の時間変化の確率を示したスキーマを記述することが必要であった．そこで本研究課題では，問題領域の知識を前提とせずに，機械学習を用いて訓練データからオブジェクトの属性の時間変化の大きさを反映したモデルを自動的に構築することを試みた．

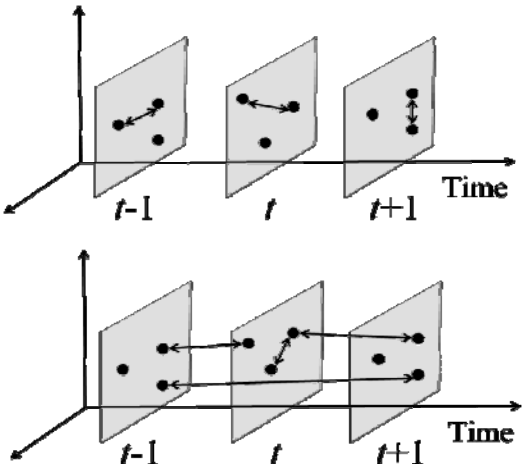

図 1 従来のリンク予測（上）と時間横断的リンク予測（下）

オブジェクト識別の問題は，対象となるデータや文書をノード，そこに記述されているオブジェクトの同一関係をリンクとみなすことで，リンク予測問題の一例と考えることができる．これまで，リンクがあるか無いかが既知のデータのペアを訓練集合として，未知なデータ間のリンクを予測するモデルを学習する方式が，様々な機械学習研究者によって提案されてきた．従来は，データは定常的であると仮定して，リンク予測が行われてきた．時間変化するデータに対してリンク予測を行う試みにおいても，同一の時点におけるデータ間にのみリンクが許されていた．一方，本研究では，２つの異なる時点におけるデータの間の時間横断的（cross-temporal）なリンクも予測の対象として扱う（図 1）．これにより，時間的に大きく離れたデータ間の同一性も判定することが可能となる．このようなリンクの予測は，従来のリンク予測の研究では扱われておらず，新しい種類の問題を提案しているといえる．

このような，時間横断的なリンク予測問題を機械学習を用いて解決するために，我々は，Vert and Yamanishi によって提案された次元削減に基づくリンク予測の方式を時間を考慮した場合に拡張した．次元削減のアプローチを採用した理由は，時間変化するデータを扱う場合，次元削減を行うことによって，異なる時点での異なる特徴が低次元空間で共通の特徴に写像されることで，表面的な特徴だけからでは同一性が分からないデータ間においても，同一性を推測することが可能となると考えたからである．

高次元のデータから，機械学習によってリンクの予測に有効な少数の特徴を抽出し，それらの特徴だけからなる比較的低次元の特徴空間においてデータ間の距離を測ることで，精度の良いリンク予測を行うことが期待

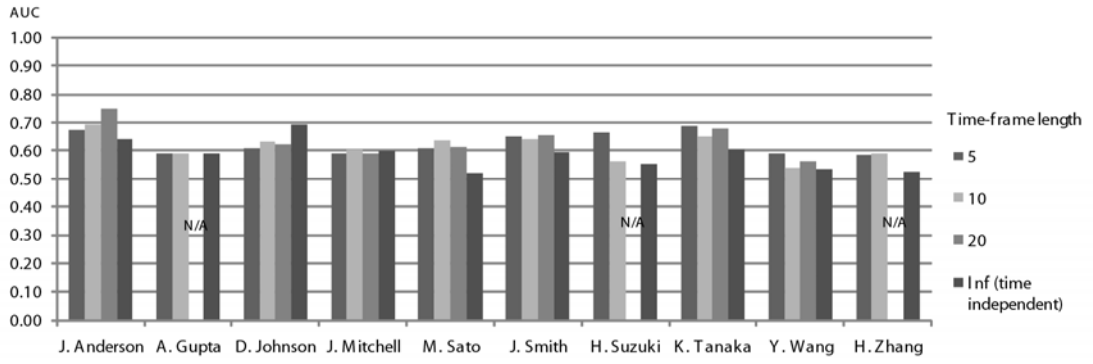

図 2 人物の同一性判定の実験結果

できる．Vert and Yamanishi によって提案された方法では，既にリンクが存在することが分かっているデータのペアを訓練集合として用い，それらのデータ間の距離が射影先で近くなるような低次元射影を学習する．リンクを予測する際には，リンクがあるか不明な 2 つのデータを学習によって得られた射影行列を用いて低次元空間に射影し，それらのデータが射影先で十分に近ければリンクがある，そうでなければリンクがないと判定する．射影行列の学習は，最適化問題として定式化され，最終的には一つの一般化固有値問題を解くことに帰着される．

時間変化するデータを扱う場合，リンク予測に有効な特徴も時間とともに変化すると考えられる．ここで，時刻が複数の時間区間に分けられているとする．そして，各時間区間で異なる射影行列を用いて，異なる時間区間のデータを共通の低次元空間に射影する．学習の際にはリンクのあるデータ同士は元の特徴空間での距離や時間的な距離にかかわらず，埋め込み先では距離が近くなるようにする．

上記の方法では，複数の時間区間にデータ集合を分割するため，各時間区間ではデータが少なくなり，オーバーフィッティングの可能性が大きくなるという問題が存在する．また，データが全くない区間では射影行列の学習ができないとう問題も生じる．一方，予測に有効な特徴は，隣接する区間では大きく変化しないことが想定される．そこで，複数の関連するタスクに関する学習を同時に行うマルチタスク学習の考え方を用いて，隣接する時間区間の射影行列は似ているようにするという要求を学習の際に課し，複数の行列の学習を同時に行うことにする．この最適化問題も，一つの一般化固有値問題に帰着することができ，それを解くだけで，全ての射影行列を一度に求めることができる．この提案手法を時間横断的局所性保存射影（Cross-temporal Locality Preserving Projections, CT-LPP）と呼ぶ．

４．研究成果

図 2 に人物の同一性判定問題における実験結果を示す．多くの例において，時間変化を考慮した提案手法が，時間変化を考慮しないベースライン（従来の LPP，一番右側の棒）に比べて，AUC（予測性能の良さを表す指標）で優っていることが示されている．

この研究成果は，データマイニング分野の主要な国際会議である 11th IEEE International Conference on Data Mining (ICDM 2011)に論文が採録され，発表を行った．この論文の成果をまとめると以下になる．時間変化するオブジェクトの同一性判定問題を，時間的に離れたデータ間の関係を予測する時間横断的リンク予測問題として定式化した．予測に有効な特徴は時間とともに変化するため，異なる時点でのデータを共通の低次元の特徴空間に射影して比較することで，精度の良い予測を行う方式の研究を行った．次元削減に基づくリンク予測の方法を拡張し，時間に依存した射影行列を学習することで，時間横断的リンク予測問題を精度良く解く方式を提案した．提案手法ではマルチタスク学習の考えを用いて，複数の時間区間の射影行列を少ない訓練データから効果的に学習する．この学習問題は行列の一般化固有値問題に帰着することで一度に求めることができる．

また，本研究課題においては，時間変化するオブジェクト情報の収集と管理に関して，オブジェクトやイベントの時間的な重要度を時間横断的なリンク解析を行うことで計算する方式，イベントの空間的曖昧性を Web における人名や地名との共起確率に基づき解決する方式，イベント名などの恣意的に名前づけられるオブジェクトの同一性を精度良く判定する手法，オブジェクト間の関係の時間変化を特定し可視化する手法，Web に日付とともに記載されている予定・予測などの明示的な将来の情報や，暗示的な将来情報を抽出する手法，人物やイベントなどの歴史的なオブジェクトの一般性及び専門性をリンク構造および時空間詳細度に基づき推定す

る手法などについても研究を行った．

今後もWebへの情報の蓄積は続いていくと予想され，Web の歴史が長くなるにつれ，相対的に現在から見て過去の情報も増加していくと考えられる．実際，インターネットアーカイブのように，積極的にWeb上の過去の情報を保存しておこうという取り組みもある．また，Web には現在や過去の事実のみならず，将来に関する予定や予測に関する情報も数多く存在している．このように時間的に広がりのある情報を効率よく扱う上で，本研究で示したような時間変化するオブジェクトを取り扱う技術の重要性は一層増していくと考えられる．

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高橋 侑久，大島 裕明，山本光穂，岩崎弘利，小山 聡，田中 克己：インパクトを考慮した歴史エンティティの重要度計算手法，情報処理学会論文誌，Vol. 52，No. 12，pp. 3542-3557，2011.（査読有）

②高橋 侑久，大島 裕明，山本光穂，岩崎弘利，小山 聡，田中 克己：Wikipedia リンク構造を用いた歴史エンティティの重要度計算，日本データベース学会論文誌，Vol. 10，No. 1，pp. 25-30，2011.（査読有）

③高橋 良平，小山 聡，田中 克己：恣意的に名前付けされたオブジェクトの識別手法，日本データベース学会論文誌，Vol. 8，No. 1，pp. 5-10，2009.（査読有）

〔学会発表〕（計10件）

①Satoshi Oyama，Kohei Hayashi，Hisashi Kashima：Cross-temporal Link Prediction，11th IEEE International Conference on Data Mining（ICDM 2011），Marriott Pinnacle Downtown（Canada），2011年12月13日.

②Yuku Takahashi，Hiroaki Ohshima，Mitsuo Yamamoto，Hirotoshi Iwasaki，Satoshi Oyama，Katsumi Tanaka：Evaluating Significance of Historical Entities Based on Tempo-Spatial Impacts Analysis Using Wikipedia Link Structure，22nd ACM Conference on Hypertext and Hypermedia（Hypertext 2011），Eindhoven University of Technology（The Netherlands），2011年6月8日.

③ Mitsuo Yamamoto，Yuku Takahashi，Hirotoshi Iwasaki，Satoshi Oyama，Hiroaki Ohshima，Katsumi Tanaka：Extraction and Geographical Navigation of Important Historical Events in the Web，10th International Symposium on Web and Wireless Geographical Information Systems（W2GIS 2011），京都大学（京都市），2011年3月3日.

④高橋 侑久，大島 裕明，山本 光穂，岩崎弘利，小山 聡，田中 克己：Wikipedia リンク構造を用いた歴史エンティティの重要度計算，第3回データ工学と情報マネジメントに関するフォーラム(DEIM 2011)，ラフォーレ修善寺（静岡県），2011年2月27日.

⑤小山 聡，鹿島 久嗣，時間横断的同一性判定のための機械学習方式，第13回情報論的学習理論ワークショップ(IBIS 2010)，東京大学（東京都），2010年11月05日.

⑥高橋 侑久，大島 裕明，小山 聡，田中克己：リンク構造分析と時空間詳細度制御に基づくイベント情報の一般性・専門性発見と提示，平成22年度情報処理学会関西支部支部大会，大阪大学中之島センター（大阪市） 2010年9月22日.

⑦ Ryohei Takahashi，Satoshi Oyama，Hiroaki Ohshima，Katsumi Tanaka：Evaluating Truthfulness of Modifiers Attached to Web Entity Names，11th International Conference on Web-Age Information Management（WAIM 2010），Minjiangyuan International Hotel（China），2010年7月17日.

⑧金澤 健介，Adam Jatowt，小山 聡，田中克己：Web からの明示的・暗示的な将来情報の抽出，第2回データ工学と情報マネジメントに関するフォーラム（DEIM 2010），淡路夢舞台国際会議場（兵庫県）2010年3月2日.

⑨Makoto P. Kato，Hiroaki Ohshima，Satoshi Oyama and Katsumi Tanaka：Query by Analogical Example：Relational Search Using Web Search Engine Indices，18th ACM Conference on Information and Knowledge Management（CIKM 2009），Asia World-Expo（中国），2009年11月3日.

⑩Hiroaki Ohshima，Adam Jatowt，Satoshi Oyama，Katsumi Tanaka：Seeing Past Rivals：Visualizing Evolution of Coordinate Terms over Time，10th International Conference on Web Information Systems Engineering（WISE 2009），Poznań University of Economics（Poland），2009年10月5日.

６．研究組織

(1)研究代表者

小山 聡（OYAMA SATOSHI）

北海道大学・大学院情報科学研究科・准教授

研究者番号：30346100