様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 6 月 2 日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：19700091
研究課題名（和文）Web からの履歴情報の発見とその呈示方式の研究

研究課題名（英文）Discovering and Presenting History Information on the Web

研究代表者
小山 聡（OYAMA SATOSHI）
京都大学・大学院情報学研究科・助教
研究者番号：30346100

研究成果の概要：実世界に存在するオブジェクトに関する多くの情報は，時間に関連付けられる場合が多い．例えば，人物の行動や発言，企業の製品発表などのイベントには，そのイベントが発生した時点がある．また，検索を行うユーザの側にとっても，それがいつの時点での情報なのかを確認することは重要であり，履歴の形で情報を整理してユーザに呈示することで，閲覧の支援が可能になると考えている．そこで，Web ページから，履歴情報を日付とイベントの組の形で抽出をし，抽出結果を年表のように時系列で呈示する方式を研究した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,800,000	0	1,800,000
２００８年度	1,500,000	450,000	1,950,000
総　計	3,300,000	450,000	3,750,000

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 履歴情報 情報抽出 機械学習 クラスタリング 確率モデル

１．研究開始当初の背景

Web は実世界の人物や企業，組織などに関する情報を得る際に重要な情報源となっている．例えば，Web を用いて研究者の研究歴などの情報や，投資先の企業に関する情報を得ることはしばしば行われる．ところが，これらの情報は多くのページに分散しており，しかも同姓同名のような名前の曖昧性があるため，目的の人物や企業の情報だけを網羅的に収集することは困難な作業である．そこで近年，人物等のオブジェクトに関する情報検索支援を目的とした研究が進められている．これらの研究の基本的なアプローチは，人名等のオブジェクト名を問い合わせとした検索結果を，同一人物（オブジェクト）に関するページが同じクラスタに属するようにクラスタリングする（オブジェクト識別やレコード同定等と呼ばれる）ことで，閲覧を容易にしようというものである．我々も，Web の構造情報とプロファイル抽出を用いる手法や訓練例からの学習を用いる手法により，オブジェクト識別の精度を向上させる研究を行ってきた．これらの従来研究を踏まえ，Web におけるオブジェクト検索の現状において，解決しなければならない問題として以下があると考える．

（１）不十分なオブジェクト識別精度

人手で設計したルールや訓練例からの学習

を用いたとしても，Web におけるオブジェクト識別精度は，文献データベース等を対象とした場合に比べて，まだまだ不十分な段階にある．これは，Web では情報が不均質であり，問い合わせのオブジェクト名を含んでいるページ内であっても，無関係な情報が混在している等の理由がある．

（2）検索結果の閲覧の困難

たとえ同じオブジェクトに関するページがクラスタリングされたとしても，多くのページに関連する情報が分散しており，逆にページ内には検索対象と無関係の情報も混在している．これらのページを一つ一つ確認していくのは依然として労力のかかる作業である．また，情報が上書きで更新される管理されたデータベースに対して，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社長の名前や売上高などは，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．

２．研究の目的

本研究課題では，これらの問題を解決するために，オブジェクトの履歴に着目する手法を研究する．実世界に存在するオブジェクトに関する多くの情報は，時間に関連付けられる場合が多い．例えば，人物の行動や発言，企業の製品発表などのイベントには，そのイベントが発生した時点がある．また，検索を行うユーザの側にとっても，それがいつの時点での情報なのかを確認することは重要であり，履歴の形で情報を整理してユーザに呈示することで，閲覧の支援を行うことを目的とする．

そのため，Web ページから，履歴情報を日付とイベントの組の形で抽出をし，抽出結果を年表のように時系列で呈示する方式を開発する．また，時間変化するオブジェクトの識別精度を向上させる方式を研究する．これに関連して，特に以下の 4 つの項目に注力して研究を行う．

（1）日付表現抽出・正規化ルールの設計

履歴情報の抽出には，ページ中に現れる日付表現の抽出・正規化を行うルールが必要である．抽出ルールは，文書中に現れる日付に対応する文字列を特定するものであり，正規化ルールは，例えば，10 月 23 日という年が省略された日付を 20061023 と一意に特定できる標準形に対応させるものである． このようなルールを用いることで，高い精度で日付表現抽出・正規化を可能とすることを目指す．

（2）ページ内からの関連箇所の特定

対象のオブジェクト名を含むページであっても．実際には，同じページ内に複数のオブジェクトの情報が混在していたり，対象オブジェクトとは無関係な情報が含まれたりする場合が多い．そこで，入力されたオブジェクト名に対応する情報だけを選択して抽出する方式を研究する．例えば，ページ内から抽出した他のオブジェクト名などを特徴として用いることで，対象オブジェクトに関連する箇所の特定精度を向上させることを試みる．また，Web ページの構造を利用することで，より正確に関連箇所を特定する方式の実現を目指す．

（3）時間変化するオブジェクトの識別

人物や企業などの実世界のオブジェクトの多くは，その属性（所属や連絡先，代表者や所在地など）が時間とともに変化する．Web 上には時期の異なる情報が混在しているため，これらの情報が誤って別オブジェクトのものであると認識されてしまう可能性がある．精度の高いオブジェクト識別を行うためには，オブジェクトの時間変化の可能性を考慮する必要がある．オブジェクトの属性値の時間変化を考慮した，精度の高いオブジェクト識別方式の実現を目指す．

（4）履歴情報の集約と呈示

抽出した履歴情報を，年表形式で時間軸上に整理して呈示する方式を研究する．年表形式での表示により，ユーザがある検索対象の時間的な変遷を容易に把握できるようにすることを目指す．

３．研究の方法

（1）日付表現抽出・正規化ルールの設計

日付表現抽出ルールは，正規表現を用いて記述した．対象とした日付表現は，年月日が揃った完全なもの（例えば 2005 年 4 月 13 日）のみならず，不完全なものも対象にした．不完全な日付表現には，年が省略されたもの（5 月 3 日）やある時点からの相対的な日付を表すもの（2 年前）などが含まれる．日付表現の正規化は，抽出された日付表現を 8 桁の数字（20050413 など）に変換するものである．不完全な日付表現に対しては，情報の補完が必要となる．補完ルールとしては，不完全な年や月を補完するものや，相対的な日付を基準となる日付を基に絶対的な日付に変換するものなどがある．

（2）ページ内からの関連箇所の特定

1 つのページ内に複数のオブジェクトに関する情報が含まれる場合には，検索対象のオブジェクトに関連する部分を特定し，その部分だけから履歴を抽出する必要がある．そこで，ページの各部分が，対象オブジェクトの記述であるか否かを判定する分類器を構築した．具体的には，HTML におけるブロックレベル要素に着目し，HTML の階層構造を基にテキストの文脈を特定した．その後，分類対象となるテキストに文脈を加えたものから特徴ベクトルを生成した．特徴としては，「4 桁の数値」

や「検索対象以外の人名」といった集約された特徴を用いた．これは，訓練データの中に現われる特定の日付表現や人名に学習された分類器が依存しないようにするためである．

（3）時間変化するオブジェクトの識別

上述の履歴抽出の手法では，例えば同姓同名の人物の履歴は区別せずに扱われるが，これを異なる人物毎にクラスタリングする手法を開発した．オブジェクトに時間変化に対応するため，一定期間にオブジェクトの属性が変化する確率を用いてオブジェクト識別を行う方式の研究を行った．例えば，1 年間に人物の所属が変わる確率や，企業の規模がある大きさで変動する確率である．

そこで，また，これらの確率モデルに基づく，オブジェクトの属性の時間変化を考慮したクラスタリングアルゴリズムを開発した．

（4）履歴情報の集約と呈示

ページから抽出した日付表現および文章を年表の形に成型して表示する方式を研究した．年表の精度を向上ため，抽出した文章の文末表現や，日付表現の後の助詞などを用いてフィルタリングを行う方式を開発した．

４．研究成果

（1）日付表現抽出・正規化ルールの設計

設計したルールを実際の Web ページに対して適用して評価実験を行い，有効性と問題点を検証した．5 つの人名に対し各 20 ページから日付表現を抽出したところ，適合率は 93.3%，再現率は 71.3%であった．適合率に比べて再現率が低いのは，例えば“2/3”といった日付なのか比率なのかが分からないといった表現の場合は，抽出を行わないという適合率を重視した設計になっているからである．これらは，候補の周辺の文章を参考にすることで，抽出できる可能性がある．例えば，日付表現は曜日の共起することが多いため，曖昧な表現と曜日が共起すれば，日付表現とみなすといったことが考えられる．

（2）ページ内からの関連箇所の特定

ページ内から指定されたオブジェクト名に対応する情報だけを選択して抽出する方式の実験を行った．対象オブジェクトに関連する箇所を人手でアノテーションした訓練集合から，関連箇所を抽出するルールを決定木学習および SVM を用いて獲得した．その際，文書構造を考慮した文脈および特徴の集約を行うことで，抽出精度を向上させることを可能にした．

（3）時間変化するオブジェクトの識別

属性の時間変化の確率を表す具体的なモデルを人物と企業を対象オブジェクトとして構築することを試みた．さらに，人物および企業のオブジェクト識別問題で実験を行い，提案手法の識別精度向上への寄与を評価した．属性値の時間変化を考慮することで，オブジェクト識別精度が向上することが確認された．

（4）履歴情報の集約と呈示

我々の提案した年表形式でのオブジェクト情報の呈示は，オブジェクトレベル検索の新たな形式の一つと考えられる．オブジェクトレベル検索は，現在の一般の検索エンジンのようにページ単位で結果を呈示するのではなく，オブジェクト単位で呈示するものである．そこでは，関連ページからドメインのスキーマにしたがってオブジェクトに関する属性値（例えば文献オブジェクトであれば著者やタイトル）を抽出してユーザに呈示するものなどが提案されている．

著者の知る限り，オブジェクトレベル検索エンジンの研究において，オブジェクトの時間変化に着目した研究はこれまでになかった．例えば，文献をオブジェクトにした場合は，出版年という属性の値は不変であるが，著者をオブジェクトとした場合，所属機関という属性の値は有効時間がある．オブジェクトの時間変化を考慮しないと，同一オブジェクトの情報が異なるオブジェクトのものと混同されたり，古い情報を誤って呈示したりといった可能性がある．一方，オブジェクトの現在の情報を呈示するだけでなく，過去の時点の情報も含めて年表形式で呈示することで，オブジェクトの全体像を把握することが容易になると考えられる．

本研究は，オブジェクトレベル検索に時間軸を加えるという新たな方向性を示した点に独自性があり，過去の情報のみならず将来の情報の検索や，単一オブジェクトの年表の呈示のみならずオブジェクト間の関係の時間変化の呈示といった方向への展開も行われている．

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 5 件）

①小山 聡：アイデンティティを推定する，人工知能学会誌，Vol.24，No. 4，2009.（査読無）

②小山 聡，田中 克己：オブジェクト識別におけるクラスタ数決定方式，電子情報通信学会論文誌，Vol.J91-D，No.3，pp.521-530，2008.（査読有）

③大島 裕明，小山 聡，田中 克己：Web 集約質問処理のための検索エンジンの関係データベースインタフェース，情報処理学会論文誌：データベース，Vol.48，No.SIG20 (TOD36)，pp.50-60，2007.（査読有）

④服部 峻，大島 裕明，小山 聡，田中 克己：継承関係と同位関係に基づく概念階層

の Web からの抽出，日本データベース学会 Letters，Vol.6，No.2，pp.9-12，2007.（査読有）
⑤大島 裕明，小山 聡，田中 克己：EaRDB：Web 集約質問処理のためのプラットフォーム，日本データベース学会 Letters，Vol.6，No.2，pp.53-56，2007.（査読有）

〔学会発表〕（計 15 件）
① Adam Jatowt, Kensuke Kanazawa, Satoshi Oyama and Katsumi Tanaka: Supporting Analysis of Future-related Information in News Archives and the Web, In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2009)*, Austin, TX, USA, June 2009 (to appear).（査読有）
②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，第 1 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 2009），静岡県掛川市，2009 年 3 月 8 日．（査読無）
③Masashi Yamaguchi, Hiroaki Ohshima, Satoshi Oyama and Katsumi Tanaka: Unsupervised Discovery of Coordinate Terms for Multiple Aspects from Search Engine Query Logs, In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2008)*, pp.249-257, Sydney, Australia, December 11, 2008.（査読有）
④高橋 良平，小山 聡，田中 克己：境界が曖昧なオブジェクトの識別のための属性時空間分布を用いたクラスタリング手法，平成 20 年度情報処理学会関西支部大会，京都府京都市，2008 年 10 月 24 日．（査読無）
⑤Satoshi Oyama and Katsumi Tanaka: How Many Objects?: Determining the Number of Clusters with a Skewed Distribution, In *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI 2008)*, pp. 771-772, Patras, Greece, July 23, 2008.（査読有）
⑥Satoshi Oyama, Kenichi Shirasuna and Katsumi Tanaka: Identification of Time-Varying Objects on the Web, In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008)*, pp.285-294, Pittsburgh, PA, USA, June 19, 2008.（査読有）
⑦小山聡，白砂健一，田中克己：属性値が時間変化するオブジェクトを識別する確率モデル，第 22 回人工知能学会全国大会，北海道旭川市，2008 年 6 月 11 日．（査読無）
⑧Hiroaki Ohshima, Adam Jatowt, Satoshi Oyama and Katsumi Tanaka：Visualizing Changes in Coordinate Terms over Time: An Example of Mining Repositories of Temporal Data through their Search Interfaces, In *Proceedings of the International Workshop on Information-explosion and Next Generation Search (INGS 2008)*, pp.61-68, Shenyang, China, April 26, 2008.（査読有）
⑨白砂 健一，小山 聡，田中 克己：オブジェクト検索における属性値の時間変化を考慮した情報集約，データベースと Web 情報システムに関するシンポジウム（DBWeb 2007），東京都目黒区，2007 年 11 月 28 日．（査読有）
⑩小山 聡，田中 克己：リンク不可例題からの距離学習とオブジェクト識別，第 10 回情報論的学習理論ワークショップ（IBIS 2007），神奈川県横浜市，2007 年 11 月 5 日．（査読無）
⑪Rui Kimura, Satoshi Oyama, Hiroyuki Toda, Katsumi Tanaka: Creating Personal Histories from the Web using Namesake Disambiguation and Event Extraction, In *Proceedings of the 7th International Conference on Web Engineering (ICWE 2007)*, Lecture Notes in Computer Science, Vol.4607, pp.400-414, Como, Italy, July 20, 2007.（査読有）
⑫白砂 健一，小山 聡，田中 克己：属性値が時間変化する Web オブジェクトの識別・検索手法の提案，夏のデータベースワークショップ(DBWS 2007)，宮城県仙台市，2007 年 7 月 4 日．（査読無）
⑬大島 裕明，小山 聡，田中 克己：Web 集約質問処理のための検索エンジンの関係データベースインタフェース，夏のデータベースワークショップ(DBWS 2007)，宮城県仙台市，2007 年 7 月 3 日．（査読無）
⑭服部 峻，大島 裕明，小山 聡，田中 克己：語の同位関連と性質の継承関連を用いた概念階層の Web からの抽出，夏のデータベースワークショップ(DBWS 2007)，宮城県仙台市，2007 年 7 月 2 日．（査読無）
⑮小山 聡，田中 克己：オブジェクト識別におけるクラスタ数決定方式，第 21 回人工知能学会全国大会，宮崎県宮崎市，2007 年 6 月 20 日．（査読無）

〔図書〕（計 1 件）
①Satoshi Oyama and Katsumi Tanaka: Distance Metric Learning from Cannot-be-linked Example Pairs, with Application to Name Disambiguation, In *Sugato Basu, Ian Davidson and Kiri Wagstaff Eds., Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapter 15, pp. 357-374, Chapman & Hall/CRC Press, 2008.

６．研究組織

(1)研究代表者

小山 聡 (OYAMA SATOSHI)
京都大学・大学院情報学研究科・助教
研究者番号：30346100