科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２３　日現在

機関番号：　３２６０６
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００１８５

研究課題名（和文）離散データ構造からの知識発見手法を用いた公共事業入札の分析

研究課題名（英文）Analyzing bids on public projects based on knowledge discovery from discrete data structures

研究代表者
久保山　哲二（Kuboyama, Tetsuji）

学習院大学・計算機センター・教授

研究者番号：８０３０２６６０

交付決定額（研究期間全体）：（直接経費）　　3,900,000円、（間接経費）　　1,170,000円

研究成果の概要（和文）：本研究では、公共事業入札のデータの分析に必要となる新しい要素技術の開発を行った。従来は主として落札率等を用いた数値的な分析が行われてきたが、これに対して、入札業者間のつながりによって形作られる構造の時系列変化に着目した分析を行うための手法を提案した。この手法を実際にコンピューターの操作ログや、ソーシャルメディアの分析に適用し、これらのデータについては有効性を示した。また、12の地方自治体について過去のデータから公共事業入札のデータベースを作成した。

研究成果の概要（英文）：This study aims to develop new methods for analyzing structural transitions of time-series graphs representing relationships among companies in bids on local government contracts, in addition to the conventional numerical methods such as regression analysis of bid rates. The effectiveness of the proposed methods have been shown by analyzing computer operation data, and social media data as the testbeds. Also, the database of bids has been created for twelve local governments.

研究分野：　知能情報学

科研費の分科・細目：　知識発見とデータマイニング

キーワード：　データマイニング　グラフマイニング　公共事業入札

１．研究開始当初の背景

(1) 適正な競争入札と談合の判断基準として落札率が広く知られている。しかし、落札率の値や、その値を目的変数とする回帰問題を解くだけでは、必ずしも競争入札の適正性を分析できるわけではない。**図１**に、ある自治体の過去 5 年間の入札率（入札価格/予定価格）の分布を示した。なお、予定価格 5000 万円以上の競争入札対象事業数 1413 件、入札数 22904 件のみを抽出しており、図のグレーの部分は落札数、黒の部分はそれ以外の入札数を表している。図において、入札率の分布に予定価格をあらかじめ知っていたことをうかがわせる壁が現れている（その他の落札率 0.9 等に現れるスパイクは、最低価格に相当する部分である）。このような観察から業者の入札行動の分析には落札率のみでは不十分であることが伺える。

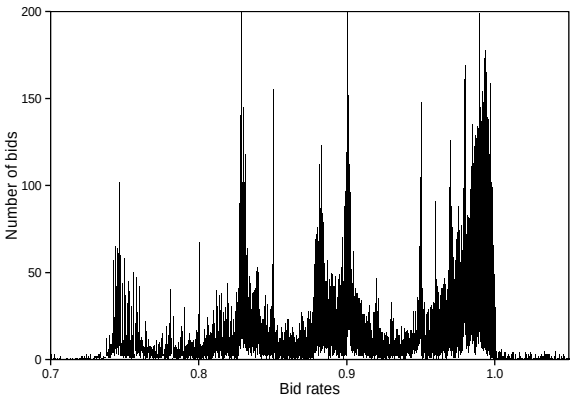

図 1 入札率と落札率の分布

(2) 代表研究者の久保山は、これまで文字列、木構造、グラフ構造等の離散データ構造の類似度設計および類似度を利用した機械学習の研究を行ってきた。また、分担者の福元は、選挙関連のデータを始めとする様々な実データを対象にした実証実験研究を行ってきた。

２．研究の目的

本研究では、膨大な入札データに潜む業者の関係性を抽出し、分析することにより、新しい入札データ分析のための要素技術を開発することを目的とする。公共事業入札のような社会的な事象を扱う場合には、いかにして推論を行ったのか説得力をもって説明できるモデルを提示することが要請される。本研究では主として次の 3 点を目的として研究を進めた。

(1) 同じ入札に参加した業者間を関連付けることにより、業者間の関係を時系列のグラフ構造として表現する。このグラフ構造に、研究代表者が継続的に開発している離散データ構造からの知識発見手法を要素技術として用い、強い結びつきを持つ業者コミュニティの候補となるグラフ構造を発見する方法を提案する。

(2) 落札率に加えて、適正な入札が行われているか否かを判定するための指標を探索する。

(3) 入札データベースの構築および実証的分析による提案手法の評価を行う。

３．研究の方法

(1) 従来の計量経済学の分野で用いられてきた談合分析の手法は、その殆どが計量ベクトルを入力とする回帰モデルの構築によるものである。機械学習の観点からは、回帰問題は教師付き学習の一種である。入札データに対しては、明確に適正でないとわかっている入札データが殆ど入手できないために、落札率を適正性の有無を表す教師データの代替として用いてきた。本研究では、入札業者のなすコミュニティ構造に着目し、構造変化を抽出する手法の開発を中心に、この問題に取り組んだ。

(2) 入札関連データの収集: 本研究の分析対象となる入札に関連したデータの収集を行う。自治体は「公共工事の入札及び契約の適正化に関する法律」により、入札者・入札金額、落札者・落札金額等の情報公開が義務付けられており、この法令に基づき開示された情報を収集する。「近年、談合が明らかになった自治体」、および「Web で公開されており、かつ、クローラーにより自動収集可能な自治体」を優先的に取得した。自治体の首長が変わった時期や談合事件の発生した時期をまたがった入札情報が Web 上で公開されていない場合には、情報開示請求を行ない、OCR や手作業によるデータ洗浄により電子的なデータを作成した。

４．研究成果

(1) 業者間の結びつきの時系列変化を検出するための要素技術として、グラフ構造の時系列変化を分析するための手法を開発した。時系列グラフの変化分析には主として(i)グラフ編集距離による方法、(ii)木の編集距離による方法、(iii)グラフカーネルによる方法を開発した。(i)グラフ編集距離による方法では、2 つのグラフ構造間の頂点と辺の差分を計算して、構造の違いを抽出した。この問題は、一般にはグラフの同型判定問題を含むため、計算コストが高いが、頂点の一意性を仮定できるグラフについては、高速にグラフ編集距離を計算できる。この手法を、SNS 上の書込の共起語から構成したワードネットワークの時系列分析に適用し、その有効性を示した（主に文献⑮）。(ii)木の編集距離による方法では、階層的コミュニティ抽出手法により、時系列ネットワークを時系列の木構造として捉え、

編集距離およびノード間の対応マッチングを用いて時系列コミュニティ構造の遷移を推定した（主に文献④）。また、同様に SNS から構成したワードネットワークに本手法を適用し、トピックの遷移を抽出した。(iii)グラフカーネルによる方法では、グラフ構造間の類似度をグラフカーネルにより計算し、カーネル PCA によってグラフ構造間の相対的な関係を低次元空間への埋め込みによって可視化およびクラスタリングした。また、PC の操作ログからのタスク遷移推論にこの手法を適用し、その有効性を示した（主に文献⑰）。**図 2** に処理の流れを示した。

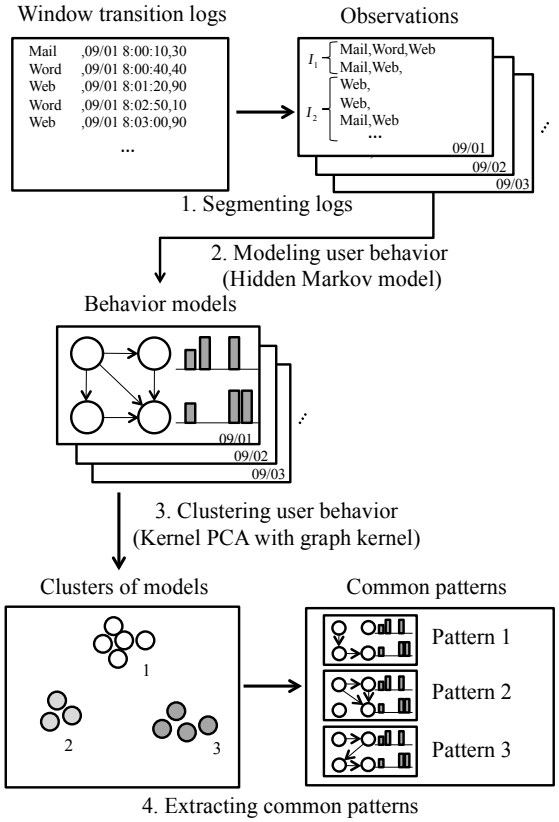

図 2 時系列のグラフ構造遷移抽出
（コンピューターの操作ログからの行動抽出による例）

(2) 全国の自治体を対象に、Web クローリングと情報開示請求により、入札データを収集した。結果、23 府県 1 市について各々過去 2～5 年間のデータが得られた。殆どは紙媒体であり、そのフォーマットも様々であるため、まず、データ洗浄と標準化を行い、リレーショナルデータベースを構築した。データベースの構築にあたっては、自治体ごとに様式の異なる入札データを統一的に扱えるように、データスキーマを設計し、予定価格、最低制限価格、調査基準価格、落札価格、落札者、事業結果（落札か不調か等）、入札者結果（落札か辞退か失格か等）、契約方式（一般競争入札、指名競争入札、随意契約）、発注機関、工事名、場所、調達区分（工事か委託か等）、業種（舗装か測量か等）、入札・契約年月日などのデータに透過的にアクセスできるようにした。研究期間中に 12 の自治体についてデータベース化が終了した。研究期間終了後も、継続してデータベース構築を行っており、今後、開発した要素技術を用いて入札データの分析を順次進めてゆく予定である。

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 18 件）査読有

① I. Hamada, T. Shimada, D. Nakata, K. Hirata, and T. Kuboyama. Agreement subtree mapping kernel for phylogenetic trees. In Proc. of JSAI-isAI Post-Workshop, volume 8417 of Lecture Notes in Artificial Intelligence, pp. 313–328. 2014.

② K. Shin and T. Kuboyama. A comprehensive study of tree kernels. In Proc. of JSAI-isAI Post-Workshop, volume 8417 of Lecture Notes in Artificial Intelligence, pp. 329–343, 2014.

③ T. Hashimoto, B. Chakraborty, T. Kuboyama, and Y. Shirota. Temporal aware- ness of needs after east japan great earthquake using latent semantic analysis. In 23rd European- Japanese Conference on Information Modelling and Knowledge Bases (EJC), pp.200–212, 2013. DOI:10.3233/978-1-61499-361-2-200

④ S. Higuchi, T. Kuboyama, T. Hashimoto, and K. Hirata. Exploring social context from buzz marketing site -- community mapping based on tree edit distance --. In Proc. Of IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM 2013 Workshops, pp.187–192, 2013. DOI:10.1109/PerComW.2013.6529479

⑤ M. Nakahara, S. Maruyama, T. Kuboyama, and H. Sakamoto. Scalable detection of frequent substrings by grammar-based compression. IEICE Transactions, 96-D(3):457–464, 2013. http://search.ieice.org/bin/summary.php?id=e96-d_3_457

⑥ Y. Nakamura, T. Horiike, T. Kuboyama, and H. Sakamoto. Extracting research communities by improved maximum flow algorithm. International Journal of Knowledge-Based and Intelligent Engineering Systems, 16(1):25–34, 2012. DOI:10.3233/KES-2012-0230

⑦ T. Hashimoto, T. Kuboyama, B. Chakraborty, and Y. Shirota. Discovering emerging topic about the east japan great earthquake in video sharing website. In TENCON 2012, IEEE Region 10 Conference, pp.1–6, 2012. DOI: 10.1109/TENCON.2012.6412324

⑧ T. Hashimoto, T. Kuboyama, B. Chakraborty, and Y. Shirota. Discovering topic transition about the east japan great earthquake in

dynamic social media. In Global Humanitarian Technology Conference (GHTC), 2012 IEEE, pp.259–264, 2012.
DOI: 10.1109/GHTC.2012.42

⑨ K. Shin and T. Kuboyama. Dynamic labeling and tree kernels with gap penalties. In Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on, pp.1690–1695, 2012.
DOI: 10.1109/SCIS-ISIS.2012.6505348

⑩ K. Shin, T. Kuboyama, and H. Nishijima. A new consistency-based feature selection algorithm. In 18th International Conference on Soft Computing MENDEL, pp.570–575, 2012.

⑪ Y. Yamamoto, K. Hirata, and T. Kuboyama. On computing tractable variations of unordered tree edit distance with network algorithms. In JSAI-isAI Workshops, volume 7258 of Lecture Notes in Artificial Intelligence, pages 211–223, 2012.
DOI:10.1007/978-3-642-32090-3_19

⑫ 木村大翼, 久保山哲二, 渋谷哲朗, 鹿島久嗣. 部分パスに基づいた木カーネル. 人工知能学会論文誌, 26(3):473–482, 2011.
https://www.jstage.jst.go.jp/article/tjsai/26/3/26_3_473/_pdf

⑬ T. Hashimoto, T. Kuboyama, and Y. Shirota. Detecting unexpected correlation between a current topic and products from buzz marketing sites. In Proc. 7th International Workshop on Databases in Networked Information Systems (DNIS), volume 7108 of Lecture Notes in Computer Science, pp.147–161, 2011.
DOI: 10.1007/978-3-642-25731-5_13

⑭ D. Kimura, T. Kuboyama, T. Shibuya, and H. Kashima. A subpath kernel for rooted unordered trees. In 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), volume 6634 of Lecture Notes in Computer Science, pp.62–74, 2011.
DOI:10.1007/978-3-642-20841-6_6

⑮ T. Kuboyama, T. Hashimoto, and Y. Shirota. Consumer behavior analysis from buzz marketing sites over time series concept graphs. In Knowledge-Based and Intelligent Information and Engineering Systems, 13th International Conference (KES), volume 5712 of Lecture Notes in Computer Science, pp.73–83, 2011.
DOI: 10.1007/978-3-642-23863-5_8

⑯ M. Nakahara, S. Maruyama, T. Kuboyama, and H. Sakamoto. Scalable detection of frequent substrings by grammar-based compression. In Proc. 14th International Conference on Discovery Science(DS2011), volume 6926 of Lecture Notes in Computer Science, pp.236–246, 2011.
DOI:10.1007/978-3-642-24477-3_20

⑰ R. Saito, T. Kuboyama, Y. Yamakawa, and H. Yasuda. Understanding user behavior through summarization of window transition logs. In Proc. 7th International Workshop on Databases in Networked Information Systems (DNIS), volume 7108 of Lecture Notes in Computer Science, pp. 162–178, 2011.
DOI:10.1007/978-3-642-25731-5_14

⑱ K. Shin, M. Cuturi, and T. Kuboyama. Mapping kernels for trees. In Machine Learning, Proceedings of 28th International Conference (ICML 2011), pp 961–968, 2011.
DOI:10.1145/1390156.1390275

〔学会発表〕（計 10 件）

① 橋本隆子，D.Shepard，久保山哲二．ソーシャルメディアにおけるバーストパターンの共起に 基づく新概念抽出．人工知能基本問題研究会，91:47-52，Nov 2013．

② 久保山哲二．公共事業入札データからのコミュニティ抽出の試み．人工知能基本問題研究会 90，Jul 2013．

③ 申吉浩，久保山哲二．木カーネルの選び方．人工知能基本問題研究会，89:61-66，Feb 2013．

④ 鈴木一寛，久保山哲二，安田浩．ネットワークの機能クラスタリングを用いた PC 操作ログ分析．人工 知能基本問題研究会，89:55-60，Feb 2013．

⑤ 久保山哲二，申吉浩．集合被覆問題の解法を用いた特徴選択．人工知能基本問題研究会，88:81-84，Jan 2013．

⑥ 樋口鐘一，橋本隆子，久保山哲二，平田耕一．バズマーケティングサイトからのコミュニティマッピングの抽出.人工知能基本問題研究会，88:133-138，Jan 2013．

⑦ 松本美玲，山抱由依，久保山哲二，坂本比呂志．大規模グラフから抽出したコミュニティの階層化．人工知能基本問題研究会，86:67-69，Aug 2012．

⑧ 申吉浩，久保山哲二．A new consistency-based feature selection algorithm．人工知能基本問題研究会，86:1-6，Aug 2012．

⑨ 齋藤良平，久保山哲二，山川裕大，安田浩．カーネル PCA を用いたユーザー行動モデルのクラスタリング．人工知能基本問題研究会，84:1-6，Dec 2011．

⑩ 齋藤良平，山川裕大，久保山哲二，安田浩．作業ウィンドウの時系列ログデータからの業務遷移パターン抽出．人工知能基本問題研究会 82:13-18，2011．

〔図書〕（計 1 件）

① H. Sakamoto and T. Kuboyama. Book Chapter: Pattern extraction from graphs and beyond. In

Multimedia Services in Intelligent Environments, pages 115–145. Springer, 2013.

６．研究組織

(1)研究代表者
久保山 哲二（KUBOYAMA, Tetsuji）
学習院大学・計算機センター・教授
研究者番号： 80302660

(2)研究分担者
福元 健太郎（FUKUMOTO, Kentaro）
学習院大学・法学部・教授
研究者番号： 50272414