様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　５月２２日現在

機関番号：８２６１６
研究種目：基盤研究(B)
研究期間：２００８　～　２０１２
課題番号：２１３００３２０
研究課題名（和文）　試験問題統計情報のデータベース化と自然言語処理技術を用いた統計的解析
研究課題名（英文）　Database creation of test-item's statistical information set and the statistical analysis using natural-language-processing technology
研究代表者
石岡 恒憲（ISHIOKA TSUNENORI）
大学入試センター・研究開発部・教授
研究者番号：８０３１１１６６

研究成果の概要（和文）：
(1)センター試験問題，全国国公私立大学入試問題に関し，所内ネットワークを介して包括的に検索表示するシステムを構築した。センター試験問題については統計情報データベースとのリンケージを図った。
(2)自然言語処理技術を用いた推薦システムの技術が本課題の要素技術の一つになるという判断のもとに,ランダムフォレストを用いた実証研究としてセンター試験の結果を用い,手法の有効性を確認した。
(3)ランダムフォレスト法が欠測を含むデータ解析手法の標準の一つになるであろうとの判断のもとに，コンテキスト情報の補完など実証的な研究や，新たな手法についての検討を行ない，多くの査読付き国際研究集会で研究成果を発表した。

研究成果の概要（英文）：
(1) We have built a comprehensive system which indicates national-center-test items linking their statistics stored in information database through our local network. Public and private university entrance examination test items can also be retrieved.
(2) Judging that the technology of recommendation system with natural-language-processing becomes one of the components of our subject, we verified the validity of the technique. As an empirical study, we use Random Forests to the scores of national center test, and announced the result to research meeting.
(3) Since convincing that Random Forests method should be recognized to one of the standards of the data-analysis technique containing missing data, we performed positive researches, such as an imputation of context information, and investigated the new technique. The results of research were presented at many refereed international research conferences.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,900,000	570,000	2,470,000
2010 年度	1,100,000	330,000	1,430,000
2011 年度	1,300,000	390,000	1,690,000
2012 年度	1,900,000	570,000	2,470,000
年度			
総　計	6,200,000	1,860,000	8,060,000

研究分野：総合領域
科研費の分科・細目：科学教育・教育工学　　教育工学
キーワード：自然言語処理，情報システム，統計数学，人工知能

１．研究開始当初の背景

平成 9 年以降の大学入試センター試験の解答データについての統計情報の整備は，研究開発部における「試験問題統計情報の整備に関する研究」（平成 14 年度～16 年度文部科学省共同研究Ｉ）において着々と行われてきた。次に我々が目指すべきものは，自然言語処理技術を用いた試験問題文そのものがもつ属性に踏み込んだデータ解析とそのデータ提供である。これら属性と成績との関係がわかれば，今後の作題において得点予測の十分な資料となることが期待される。そして，このような自然言語を取り扱うための道具立ては，現在かなり整ってきている。形態素解析については，代表的なものだけでも JUMAN（京都大学），ChaSen（茶筌，奈良先端），MeCab(和布蕪）などが知られる。構文解析についても KNT(京都大学）や CaboCha(南瓜，奈良先端）などがあり，フリーで使うことができる。加えて研究代表者は，日本で唯一の小論文自動採点システム Jess の開発者であり，本研究の分野について専門性と開発実績を有している。

２．研究の目的

試験問題の属性には

・素材文，あるいは選択肢の難読性に関するもの

・出題形式に関するもの（質問形式，選択肢数など）

・素材文のジャンル（現代文(評論・小説）/古文/漢文，会話文/説明文などの別）

に関するものがあり，これらの属性による難易度予測と，試験情報データベースとのシームレスな統合を実現するために，システムの開発やデータベースの整備を図る。またこれらの属性は，人間が判断するものでなく，可能な限り，最新の自然言語処理技術を用いて自動的に取得する。

３．研究の方法

本研究は，センター試験の採点結果を直接的に取り扱うために，メンバーは研究代表者が所属する大学入試センター研究開発部の教員のみによって組織する。大よその分担は以下の通り。

石岡（研究代表者）：統括，自然言語処理

橋本：過去の統計情報との統合，データベース

大津：XML 構造化

システム試作についての技術課題について，以下の通り実施した：

(1) 試験問題の言語処理上の属性情報の抽出（石岡・大津・橋本）

国語/英語/社会など試験問題の属性情報の項目の洗い出しを行い，それで過不足がないかの検討を行なった。

(2) 現在の試験問題情報データベースとの統合（橋本・石岡）

現在の試験問題情報データベースでは，文字列一致による試験問題の検索とそれに付随する各種統計情報の表示が可能であるが，今回の試験問題の属性情報も必要に応じ引けるようにした。

(3) マシン環境の整備（大津・石岡）

LAN 上で Web ブラウザによって統計情報を容易に検索できるように，今回，獲得された試験問題属性に関する統計量を記述した HTML ファイルを置く必要がある。またデータベースによる検索機能を一部，合わせる必要があり，これを実施した。

(4) XML 化に向けた関連調査（大津・石岡・橋本）

格納データの XML 化についても検討を加える。このために e ラーニング（e-Learning）標準規格である SCORM 1.2（スコーム 1.2）の概要について習得する。短期間で習得するために，有料の講座を受講した。

システム完了後は，各研究者の興味・関心にしたがって，データ利用に関わる統計手法の開発を行った。このフェーズでは，統計解析の専門家である桜井や，外部よりシステムの利活用に関心のある中済をメンバーに加え，本システムの成果発表を促進させた。

４．研究成果

(1) センター試験問題，全国国公私立大学入試問題に関し，所内ネットワークを介して包括的に検索表示するシステムを構築することについて。

大学入試センター試験について，解答に関わる統計情報をデータベース化し，また過去に実施した全ての法科大学院適性試験の試験問題について，統計情報をデータベース化した。

これら試験問題データベースにおいて，検索・解析のための統合環境の構築を行い，センター試験問題については統計情報データベースとのリンケージを図った。

(2) 自然言語処理技術を用いた推薦システムの技術が本課題の要素技術の一つになるという判断のもとに，近年注目されているランダムフォレスト法の適用について実証的な実験を行うことについて。ランダムフォレストを用いた実証研究としてセンター試験の結果を用い，手法の有効性を確認した。また統計関連 5 学会共同による統計関連学会連合大会（参加者約 900 名）の研究集会にてその成果を発表した。

平成 21 年度には，情報推薦の要素技術を

確立ならびに洗練させるために，日本ＯＲ学会，人工知能学会等が主催するリコメンデーションコンテストに参加し，上位入賞を果たした。
(3) ランダムフォレスト法が欠測を含むデータ解析手法の標準の一つになるであろうとの判断のもとに，コンテキスト情報の補完など実証的な研究や，新たな手法についての検討を行なうことについて。多くの査読付き国際研究集会で研究成果を発表した。うちiiWAS2012 に関して，採択論文は全て ACM International Conference Proceedings Series (ISBN: 978-1-4503-1306-3)に掲載されるが，当該論文は selected paper として IIWAS2012 Special Issue Journal: Intl J of Business Intelligence and Data Mining (IJBIDM)にボリューム拡張の上，掲載されることになった。

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１２件）
①Hideki Nagatsuka, Toshinari Kamakura, Tsuneori Ishioka, An Efficient Bayesian Estimation of Ordered Parameters of Two Exponential Distributions, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol.E-92-A, No.7, 2009, 1608-1614
②石岡 恒憲, 小論文自動採点, 小特集　学力評価の最前線, 電子情報通信学会誌, **92** (12), 2009, 1036-1040.
③石岡 恒憲, 自然言語処理技術を用いたセンター試験問題の統計的解析 ―英語および国語の試験問題を対象として―, 大学入試研究ジャーナル, **20,** 2009, 145－150.
④Sakurai, H. and Taguri, M., Test of mean difference for longitudinal data using circular block bootstrap., COMPSTAT2010 Proceedings in Computational Statistics (eds. Lechevallier, Y. and Saporta, G.), Physica-Verlag, 2010, 1581-1588.
⑤青木敏・大津起夫・竹村彰通・沼田泰英, 大学入試センター試験科目選択データの統計解析, 応用統計学, **39,** 2010**,** 71-100.
⑥大津起夫, 大学入試センター試験における科目別得点の非線形因子分析による比較, 大学入試センター紀要, **40,** 2010**,** 1－23.
⑦ Tsunenori Ishioka, Ascertaining and graphically representing the logical structure of Japanese essays, International Journal of Continuing Engineering Education and Life-Long Learning (IJCEELL), Volume 21, Issue 4, 2011, 276-288.
⑧石岡恒憲, 橋本貴充, 大津起夫, 試験問題データベース検索・解析のための統合環境の構築 ―センター試験問題データベースとのリンケージ―, 大学入試研究ジャーナル, **22**, 2012, 73-78.
⑨桜井裕仁, ブートストラップ信頼区間に基づく本追モニター調査の平均点差の検討, 大学入試センター研究紀要, **41**, 2012.
⑩ Tsunenori Ishioka, Imputation of Missing Values for Semi-supervised Data Using the Proximity in Random Forests, iiWAS 2012, ACM International Conference Proceedings Series (ISBN: 978-1-4503-1306-3), iiWAS, 2012, 319-325.
⑪ Tsunenori Ishioka, Imputation of Missing Values for Unsupervised Data Using the Proximity in Random Forests, eLmL 2013 : The Fifth International Conference on Mobile, Hybrid, and On-line Learning, ISBN: 978-1-61208-253-0, eLmL2013, 30-36.
⑫中済 光昭, 記述式課題における添削支援システム　－事例ベース推論によるアプローチー, 駒沢大学経済学論集, 44（2）, 2012, 1-30.

〔学会発表〕（計１５件）
①石岡 恒憲・橋本 貴充・大津 起夫, センター試験・英語と国語における素材文のリーダビリティと得点率についての統計的解析, 電子情報通信学会　信学技報 TL2009-1, 2009 年 6 月 18 日, 機械振興会館（芝公園）.
②石岡 恒憲, 短答式記述テストにおける自動採点　－その採点ロジックと課題についてー, 電子情報通信学会　信学技報 TL2009-2, 2009 年 6 月 18 日, 機械振興会館（芝公園）.
③石岡 恒憲, 自然言語処理技術を用いたセンター試験問題の統計的解析 ―英語および国語の試験問題を対象として―, 2009 年度統計関連学会連合大会, 2009 年 9 月 7 日, 同志社大学.
④石岡 恒憲, アメリカの大学入試制度と統計の出題, 第 6 回統計教育の方法論ワークショップ, 日本統計学会統計教育分科会・日本統計学会統計教育委員会, 2010 年 3 月 5 日, 成蹊大学.
⑤石岡 恒憲・橋本 貴充・大津 起夫, 自然言語処理技術を用いたセンター試験問題の統計的解析 ―英語および国語の試験問題を対象として―, 平成 21 年度全国大学入学選抜連絡協議会, 2009 年 5 月 21 日, 日本学術総合センター.
⑥石岡 恒憲, Random forest 法を用いた推薦システムとその評価　～ 第 1 回リコメンデーションコンテストに参加して ～, 電子情報通信学会信学技報, **110** (301), AI2010-36, 2010 年 11 月 19 日, 九州大学（福岡）
⑦石岡 恒憲, Random forest 法を用いた推薦システムとその性能評価について, 2010 年度統計関連学会連合大会, 2010 年 9 月 6 日, 早稲田大学（東京）.
⑧中済光昭, 事例ベース推論を用いたレポート添削支援システム, 情報処理学会 コンピュータと教育研究会 107回研究発表会, 2010年11月20日,

香川大学 幸町キャンパス.
⑨ Hashimoto, T. & Ueno, M., Latent Conditional Independence Test for Bayesian Network IRT, First International Workshop on Advanced Methodologies for Bayesian Networks (AMBN 2010), 2010 年 11 月 18 日, Campus Innovation Center, Tokyo
⑩橋本貴充・植野真臣, LCI Test -項目間局所従属性検出のための潜在的条件付き独立性検定-,日本テスト学会第8回大会, 2010年8月30日, 多摩大学.
⑪橋本貴充・植野真臣, 受検者の能力の影響を考慮した項目の従属関係の検出, 日本教育工学会研究報告, 2010年7月3日, 電気通信大学.
⑫大津起夫・橋本貴充, 非線形因子分析による入学試験問題の難易度比較, 日本テスト学会第8回大会, 2010年8月30日, 多摩大学.
⑬青木敏・大津起夫・竹村彰通・沼田泰英, 大学入試センター試験科目選択データの統計解析, 応用統計学会2010年度年会, 2010年5月20日, 統計数理研究所.
⑭石岡恒憲・橋本貴充・大津起夫, 試験問題データベース検索・解析のための統合環境の構築―センター試験問題データベースとのリンケージ, 第6回全国入学者選抜研究連絡協議会, 2011年5月26日, 早稲田大学.
⑮石岡恒憲, 教師なしデータおよび半教師データにおけるランダムフォレストによる欠測値補完, 2012年度 統計関連学会連合大会, 2012年09月11日, 北海道大学.

〔図書〕(計２件)
①石岡 恒憲ほか(植野真臣・永岡慶三　共編), 培風館, 論述式項目の自動採点, ｅテスティング, 2012, 269.
②石岡 恒憲ほか（植野真臣・永岡慶三・山内 祐平　共編）, ミネルヴァ書房, 自然言語処理と学習評価, 教育工学における学習評価, 教育工学選書, 2012, 224.

〔産業財産権〕
該当なし

〔その他〕
ホームページ等
http://www.rd.dnc.ac.jp/~tunenori/

６．研究組織
(1)研究代表者
石岡 恒憲（ISHIOKA TSUNENORI）
大学入試センター・研究開発部・教授
研究者番号：８０３１１１６６

(2)研究分担者
大津起夫（OTSU TATSUO）
大学入試センター・研究開発部・教授
研究者番号：１０２０３８２９

橋本 貴充（HASHIMOTO TAKAMITSU）
大学入試センター・研究開発部・助教
研究者番号：２０３９９４８９

櫻井 裕仁（SAKURAI HIROHITO）
大学入試センター・研究開発部・准教授
研究者番号：００３３３６２５

中濟 光昭（NAKAZUMI MITSUAKI）
駒澤大学・経済学部・教授
研究者番号：６０３０６９１７