様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ５ 月 ２５ 日現在

機関番号：３２７０６
研究種目：基盤研究(C)（一般）
研究期間：2016～2019
課題番号：１６Ｋ０１２６７

研究課題名（和文）言語に依存しない大規模テキストデータからの自動単語分割技術の確立

研究課題名（英文）Establishment of Automatic Word Segmentation Technology from Large-scale Text Data Independent of Language

研究代表者
鈴木　誠（Suzuki, Makoto）

湘南工科大学・工学部・教授

研究者番号：８０３３９７９６

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：本研究では、Unicodeで表現された複数の言語が混在するテキストデータを、同一のプログラムで処理する単語分割技術を構築した。この技術は、単純な状態遷移モデルに基づいた、各言語の辞書や文法知識を一切必要としない言語独立な単語分割方式である。主に、(1)処理対象となる言語の拡張、(2)適用事例の拡張、の2つの方向性を意識して研究を進めた。(1)については、日本語以外にも日本語の古典や英語、中国語、韓国語などの外国語に対しても有効であることが確認できた。(2)については、商品や施設のユーザレビューを用いて感情極性辞書を自動的に作成する手法を提案することができた。

研究成果の学術的意義や社会的意義
本研究では、対象のレビューデータをもとに感情極性辞書を自動的に作成する手法を提案することができた。感情極性辞書とは、文章に含まれる単語に対し、文中に含まれる特有の極性（ポジティブ、ネガティブ）を持つ単語が含まれているという考えに基づき、単語に対し極性値を与えた辞書である。今回は商品や施設のユーザレビュー（5段階の評価値付きのテキストデータ）を用いて、評価値に基づいて感情極性値を算出することにより、感情極性辞書を自動的に作成する手法を提案した。これにより、コンピュータが自動的にユーザレビューを収集し、ある商品や施設に特化した感情極性辞書を構成できる可能性を示唆することができた。

研究成果の概要（英文）：In this research, we constructed a word segmentation technology that processes text data that is mixed with multiple languages expressed in Unicode with the same program. This technique is a language-independent word segmentation method based on a simple state transition model that does not require any dictionary or grammatical knowledge for each language. The research proceeded mainly in two directions: (1) extension of the language to be processed and (2) extension of application cases. Regarding (1), We confirmed that it is effective not only for Japanese but also for Japanese classics and foreign languages such as English, Chinese, and Korean. Regarding (2), we were able to propose a method for automatically creating an emotional polarity dictionary using user reviews of products and facilities.

研究分野：知識発見とデータマイニング

キーワード：多言語処理　感情極性辞書

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

近年、IT 技術の進歩により計算機の性能が向上し、様々な文書が電子化され、Web 上で利用可能なテキストデータの量が年々増加している。これに伴い、最近ではビッグデータの活用が盛んに叫ばれるようになり、Web 上のテキストデータを対象とした分析も積極的に行われるようになってきた。この動きは日本国内にとどまらず、新興国も含めた世界各国で活発化してきており、言語の垣根を越え、グローバル化が急速に進行している。一方で、日本国内においても SNS が急速に普及し、一般の辞書には掲載されていない口語調の文書や絵文字なども頻繁に用いられるようになっている。このような状況の中で、大量に存在するテキストデータを自動的に処理するテキストマイニング技術に対する需要も高まってきている。このテキストマイニングの中にも、文書検索・文書要約・単語分割・文書分類・Web マイニング・マーケティングなどの様々な課題が存在する。この中で、本研究では単語を自動的に切り出す問題に焦点をあてる。

日本語や中国語のような非分かち書き言語に対して単語分割を行う場合は、あらかじめ作成された辞書を参照した形態素解析が必要である。日本語においては、従来から茶筌などの形態素解析ツールが用いられてきた。しかし、このような形態素解析ツールを用いた手法で Web 上などに存在するテキストデータを正しく単語に分割することは非常に難しい。なぜなら、このような分析で対象となるテキストデータの多くは形態素解析ツールが苦手とする口語体で書かれている上、略語や新語が続々と生み出されており、形態素解析で必要となる辞書のメンテナンスもままならないためである。そこで対象となるテキストデータ自体から単語情報を認識することができる単語分割手法が必要となる。さらに、近年のスマートフォン等のモバイル端末の爆発的な普及により、Web 上のテキストデータは世界的に増え続けている。今後も日本語だけではなく中国語のような非分かち書き言語において、単語を切り出す技術は重要になると考えている。

２．研究の目的

本研究の目的は、Unicode で表現された複数の言語が混在するテキストデータを、同一のプログラムで処理する単語分割技術を構築することである。この技術は、単純な状態遷移モデルに基づいた、各言語の辞書や文法知識を一切必要としない言語独立な単語分割方式であり、かつ口語体文書や古語も含めた世界中のあらゆる非分かち書き言語が混在している大規模なテキストデータも処理可能である。また、高性能なコンピュータではなく、一般の個人が使用する PC のレベルで動作可能である。そして、文書分類や Web マイニング、さらにはソーシャルメディア時代のマーケティング等の分野でも利用できるように発展させ、言語を問わず消費者の口コミ情報など口語調の単語や新語や流行語も発見し、マーケティングにも利用できるようにする。

３．研究の方法

本研究では、日本語や中国語のような世界中の非分かち書き言語を対象として言語固有の文法や単語の知識を一切用いることなく、対象とした言語の辞書を自動的に構成し、単語を切り出す手法を構築する。単純な状態遷移モデルを仮定し、単語分割に必要な辞書を作成する。

主に、以下の 2 つの方向性を意識して研究を進めた。

(1)処理対象となる言語の拡張

様々な言語に対して単語分割の実験を行い、本手法の高い分割性能を実証する。

(2)適用事例の拡張

マーケティングや Web マイニングなどに適用問題を拡張する。

４．研究成果

「３．研究の方法」に記載した「(1) 処理対象となる言語の拡張」と「(2) 適用事例の拡張」の二点について主に研究を進めた。ここでは、(2)について以下に説明する。

本研究では、事例研究として特定の商品や施設についてのユーザレビューの分析を行った。

今回は具体例として、ゴルフ場のレビューデータ（5 段階の評価値付きのテキストデータ）を用いた。ユーザの年齢（年代）と 5 段階評価に基づき表に示すようにクロス集計を行った。その結果を表 1 に示す。ここで、5 段階評価は、高評価なものから順に「5:Best(B)」、「4:Good(G)」、「3:Average(A)」、「2:Poor(P)」、「1:Worst(W)」と表記している。クロス集計表の各セルの数値は、一人のユーザのレビューを 1 件としてカウントした。

表１：クロス集計表

個数 / 評価	W	P	A	G	B総計	
20代	0	5	47	67	17	136
30代	6	27	223	395	101	752
40代	4	22	236	343	74	679
50代	0	8	99	168	19	294
60代	0	1	16	34	3	54
70代	0	0	0	3	0	3
null	3	11	53	81	7	155
総計	**13**	**74**	**674**	**1091**	**221**	**2073**

次に、表１のクロス集計表を用いて対応分析を行った。対応分析とは、クロス集計表などの行と列からなるデータの特徴を図示し、項目間の関係を視覚的に把握する分析手法である。その結果の一例を図１に示す。

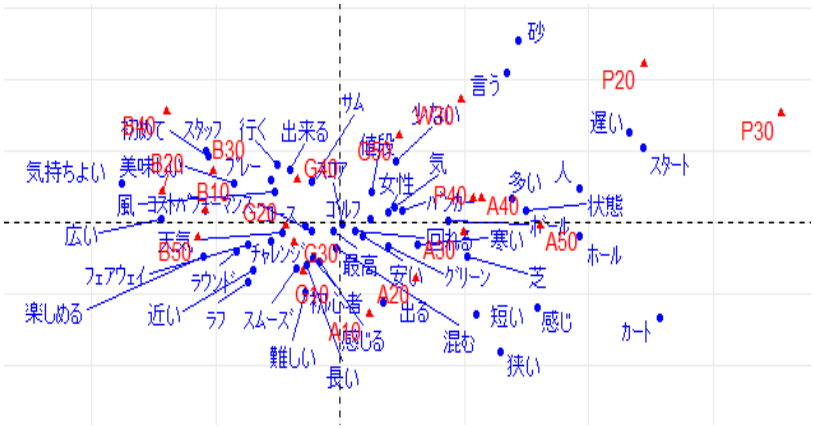

図１：対応分析の結果

そこで、対象のレビューデータを用いて感情極性辞書を自動的に作成する研究に取り組んだ。感情極性辞書とは、文章に含まれる単語に対し、文中に含まれる特有の極性（ポジティブかネガティブか）を持つ単語が含まれているという考えに基づき、単語に対し極性値を与えた辞書である。簡単な例として「うれしい」や「楽しい」などの単語は感情面でポジティブであり、「悲しい」や「辛い」という単語はネガティブであると考えられる。また、辞書によっては単にポジティブかネガティブという情報だけではなく、ポジティブおよびネガティブ度合いを数値化する場合もある。本研究は後者の立場で研究を進めた。感情極性辞書の利点として、単語レベルでポジティブかネガティブの判別ができるので直感的に感情極性の判断ができる。従来の感情極性辞書は、一般的な事柄を対象とし辞書の評価を行っているため、対象に対して評価尺度と辞書の評価尺度が一致していないことが多かった。このため、ある特定の商品や施設に対しては、本来と異なった感情極性が与えられてしまう可能性がある。そこで、本研究では特定の対象に対し、それに特化した辞書の作成が必要であると考えた。例えば「難しい」という単語は一般的にはネガティブな感情極性値が与えられる単語であるが、戦略性を求めるゴルファーのコメントデータの中に表れている回数が多く高評価なレビューの中に含まれている単語であるため、ゴルフ場においてはポジティブな感情極性値を与えて良い単語になる。このような観点から、ユーザレビューの５段階評価値がそのレビューに含まれる単語の感情極性値を決定する手法を提案した。具体的にはその計算過程に TFIDF 値を用いた。

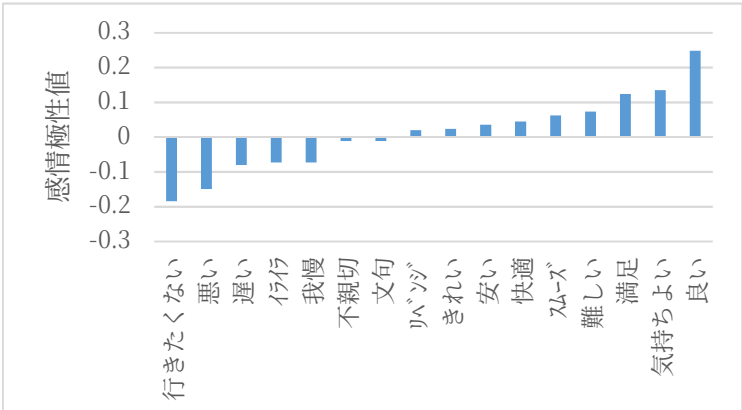

図２：感情極性値の計算結果（一部抜粋）

以上のように、ある特定の商品や施設に対するユーザレビューに対して感情極性辞書を自動的に作成する手法を提案した。しかし、今回の感情極性値の計算は、対象の文章の文脈などが考慮されていないため、文章の中でポジティブとネガティブが反転する場合は誤った感情極性値を付与してしまう可能性がある。今後は文脈も考慮した感情極性値の算出方法を検討していく必要がある。

５．主な発表論文等

〔雑誌論文〕　計15件（うち査読付論文　15件／うち国際共著　0件／うちオープンアクセス　10件）

１．著者名 安井一貴，中野修平，三川健太，後藤正幸	４．巻 Vol.28, No.2
２．論文標題 周期性とイベント効果に着目した消費者の購買行動分析モデルに関する一考察	５．発行年 2019年
３．雑誌名 経営情報学会誌	６．最初と最後の頁 pp.69-87
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 S. Nagamori, K. Mikawa, M. Goto, and T. Ogihara	４．巻 Vol.18, No.3
２．論文標題 An Analytic Model to Represent Relation between Finish Date of Job-Hunting and Time-Series Variation of Entry Tendencies	５．発行年 2019年
３．雑誌名 Industrial Engineering & Management Systems	６．最初と最後の頁 pp.292-304
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 仁ノ平 将人，三川健太，後藤正幸	４．巻 Vol.60, No.4
２．論文標題 販売履歴データに基づく中古ファッションアイテムの販売価格予測モデルに関する一考察	５．発行年 2019年
３．雑誌名 情報処理学会論文誌	６．最初と最後の頁 pp.1151-1161
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 馬賀嵩士，三川健太，後藤正幸，吉開朋弘	４．巻 Vol. J101-D, No. 7
２．論文標題 気象情報とTweetデータの統合的分析による体感気温の定量化とその需要予測への応用	５．発行年 2018年
３．雑誌名 電子情報通信学会論文誌D	６．最初と最後の頁 1037-1051
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 関口あゆみ，仁ノ平 将人，三川健太，後藤正幸	4．巻 Vol.27，No.1
2．論文標題 推定購買確率と予測評価値をバランスする意外性指標に基づく推薦システム	5．発行年 2018年
3．雑誌名 経営情報学会誌	6．最初と最後の頁 67-78
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 松嵜祐樹，三川健太，後藤正幸	4．巻 Vol.58，No.12
2．論文標題 マルコフ潜在クラスモデルに基づくECサイトにおける施策実施効果分析に関する一考察	5．発行年 2017年
3．雑誌名 情報処理学会論文誌	6．最初と最後の頁 2034-2045
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 S. Misawa, K. Mikawa, and M. Goto	4．巻 Vol.16, No.3
2．論文標題 Adaptive Prediction Method Based on Alternating Decision Forests with Considerations for Generalization Ability	5．発行年 2017年
3．雑誌名 Industrial Engineering & Management Systems	6．最初と最後の頁 384-391
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 L. Suzuki, K. Mikawa, and M. Goto	4．巻 Vol.16, No.2
2．論文標題 Multi-Valused Classification of Text Data Based on an ECOC Approach Using a Ternary Orthogonal Table	5．発行年 2017年
3．雑誌名 Industrial Engineering & Management Systems	6．最初と最後の頁 155-164
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Y. Yamamoto, K. Mikawa, and M. Goto	4．巻 Vol.16, No.2
2．論文標題 A Proposal for Classification of Document Data with Unobserved Categories Considering Latent Topics	5．発行年 2017年
3．雑誌名 Industrial Engineering & Management Systems	6．最初と最後の頁 165-174
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 湯川 輝一朗，三川健太，後藤正幸	4．巻 Vol.68, No.2
2．論文標題 データの転送制御に基づいた分散型SVMの効率的な学習手法	5．発行年 2017年
3．雑誌名 日本経営工学会論文誌	6．最初と最後の頁 86-98
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 藤原直広，三川健太，後藤正幸	4．巻 Vol.26, No.1
2．論文標題 閲覧及び購買行動を同時に表現するアスペクトモデルによる購買予測手法の提案	5．発行年 2017年
3．雑誌名 経営情報学会誌	6．最初と最後の頁 1-16
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 早川真央，三川健太，荻原大陸，後藤正幸	4．巻 Vol.58, No.5
2．論文標題 層別木と混合ワイブル分布に基づく就職活動終了時期の分析モデルの構築	5．発行年 2017年
3．雑誌名 情報処理学会論文誌	6．最初と最後の頁 1189-1206
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 T. Maga, K. Mikawa, and M. Goto	４．巻 Vol.3, No.1
２．論文標題 Data pair selection for accurate classification based on information-theoretic metric learning	５．発行年 2017年
３．雑誌名 Asian J. Management Science and Applications	６．最初と最後の頁 61-74
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Bin Xu, Naohide Yamagishi, Makoto Suzuki, Masayuki Goto	４．巻 Vol.15, No.3
２．論文標題 Language-Independent Word Acquisition Method Using a State-Transition Model	５．発行年 2016年
３．雑誌名 J. Industrial Engineering & Management Systems	６．最初と最後の頁 197-207
掲載論文のDOI（デジタルオブジェクト識別子） 10.7232/iems.2016.15.3.224	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 三川健太，後藤正幸	４．巻 Vol.66, No.4
２．論文標題 カテゴリ毎に異なる計量行列を用いた計量距離学習手法に関する一考察	５．発行年 2016年
３．雑誌名 日本経営工学会論文誌	６．最初と最後の頁 335-347
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計23件（うち招待講演　0件／うち国際学会　11件）

１．発表者名 M.Suzuki, T.Onuma, N.Katsumata and N.Yamagishi
２．発表標題 Analysis of Review Data on Educational Toys
３．学会等名 International Conference on Education（国際学会）
４．発表年 2020年

１．発表者名
勝間田 昇，山岸 直秀，鈴木 誠

２．発表標題
ゴルフ場のレビューデータを用いた感情極性辞書の作成

３．学会等名
経営情報学会春季全国研究発表大会

４．発表年
2019年

１．発表者名
小沼 拓也，山岸 直秀，鈴木 誠

２．発表標題
知育玩具に関するレビューデータの分析

３．学会等名
経営情報学会春季全国研究発表大会

４．発表年
2019年

１．発表者名
K. Mikawa

２．発表標題
Factorization Machine with Mixed Norm Regularization using ADMM

３．学会等名
20th Asia Pacific Industrial Engineering and Management System (APIEMS2019)

４．発表年
2019年

１．発表者名
Kuwata, T. Sugisaki, K. Mikawa, and M. Goto

２．発表標題
An Estimation Model of Open Price for Second-hand Fashion Items Based on Sales History Data

３．学会等名
Proc. 17th Asian Network for Quality Congress (ANQ2019)

４．発表年
2019年

１．発表者名
Y. Sakai, K. Yasui, K. Mikawa, and M. Goto

２．発表標題
An Extension of Semi-Supervised Boosting to Multiclass Classification

３．学会等名
Proc. 17th Asian Network for Quality Congress (ANQ2019)

４．発表年
2019年

１．発表者名
K. Yasui, K. Mikawa, and M. Goto

２．発表標題
An Analytical Model of Consumers Purchasing Behavior Considering the Variety of Products

３．学会等名
Proc. 17th Asian Network for Quality Congress (ANQ2019)

４．発表年
2019年

１．発表者名
T. Sugisaki, K. Mikawa, and M. Goto

２．発表標題
Factorization Machines Considering the Latent Characteristics Behind Target Data

３．学会等名
Proc. 17th Asian Network for Quality Congress (ANQ2019)

４．発表年
2019年

１．発表者名
Shuhei Nakano, Kenta Mikawa, and Masayuki Goto

２．発表標題
A Study of the Application of Canonical Correlation Forests to Text Classification

３．学会等名
The 19th Asia Pacific Industrial Engineering and Management Systems (APIEMS2018)（国際学会）

４．発表年
2018年

１．発表者名
Kenta Mikawa, Manabu Kobayashi, Masayuki Goto, and Shigeichi Hirasawa

２．発表標題
Proposal for an l1 regularized Factorization Machine

３．学会等名
The 19th Asia Pacific Industrial Engineering and Management Systems (APIEMS2018)（国際学会）

４．発表年
2018年

１．発表者名
Tomoya Sugisaki, Yuri Nishio, Kenta Mikawa, Masayuki Goto, and Takashi Sakurai

２．発表標題
A New Entry Behavior Model of Student Users on Job Board for New Graduates Considering the Interaction between Features

３．学会等名
16th Asian Network for Quality Congress（国際学会）

４．発表年
2018年

１．発表者名
安井一貴，中野修平，三川健太，後藤正幸

２．発表標題
周期性とイベント効果に着目した消費者の購買行動分析モデルに関する一考察

３．学会等名
日本経営工学会平成30年度春季研究大会予稿集

４．発表年
2018年

１．発表者名
杉崎智哉，西尾友里，三川健太，後藤正幸，桜井 崇

２．発表標題
特徴間の交互作用を考慮した学生ユーザの企業エントリー行動分析モデルに関する一考察

３．学会等名
日本経営工学会平成30年度春季研究大会予稿集

４．発表年
2018年

１．発表者名
三川健太，小林 学，後藤正幸，平澤茂一

２．発表標題
l1正則化に基づくFactorization Machineに関する一考察

３．学会等名
日本経営工学会平成30年度春季研究大会予稿集

４．発表年
2018年

１．発表者名
畠山一輝，三川健太，小林学

２．発表標題
MineCraftを用いたDQNによる構造物の自動構築の検討

３．学会等名
2018年電子情報通信学会ソサイエティ大会予稿集

４．発表年
2018年

１．発表者名
中野修平，三川健太，後藤正幸

２．発表標題
Canonical Correlation Forests におけるラベル行列のスパース性を考慮した 分類法に関する一考察

３．学会等名
情報処理学会第121回数理モデル化と問題解決研究発表会

４．発表年
2018年

１．発表者名
M.Suzuki, N.Yamagishi, K.Mikawa and M.Goto

２．発表標題
Characteristics of a Word Segmentation Method Based on a State-transition Model

３．学会等名
Proc. of Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS2017)（国際学会）

４．発表年
2017年

１．発表者名
K. Mikawa, M. Kobayashi, M. Goto, and S. Hirasawa

２．発表標題
Distance Metric Learnig using Each Category Centroid with Nuclear Norm Regularization

３．学会等名
The 2017 IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2017)（国際学会）

４．発表年
2017年

１．発表者名
Makoto Suzuki, Bin Xu, Naohide Yamagishi, Masayuki Goto

２．発表標題
Word Acquisition of Japanese Classical Literature Using State Transition Model

３．学会等名
Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS2016)（国際学会）

４．発表年
2016年

１．発表者名
Kenta Mikawa, Manabu Kobayashi, Masayuki Goto, Shigeichi Hirasawa

２．発表標題
A Study on Distance Metric Learning using Distance Structure among Category Centroids

３．学会等名
Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS2016)（国際学会）

４．発表年
2016年

１．発表者名
Yusei Yamamoto, Kenta Mikawa, Masayuki Goto

２．発表標題
A proposal of document recommendation based on topic model

３．学会等名
Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS2016)（国際学会）

４．発表年
2016年

１．発表者名
Yuki Matsuzaki, Kenta Mikawa, Masayuki Goto

２．発表標題
Modeling customer purchase behavior based on page transitions by latent class model for customer segmentation

３．学会等名
Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS2016)（国際学会）

４．発表年
2016年

１．発表者名
Qian Zhang, Haruka Yamashita, Kenta Mikawa, Masayuki Goto

２．発表標題
A study of extended RFM analysis based on PLSA model for Purchase History Data

３．学会等名
Asia Pacific Industrial Engineering and Management Systems Conference (APIEMS2016)（国際学会）

４．発表年
2016年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	三川　健太 (Mikawa Kenta) (40707733)	湘南工科大学・工学部・准教授 (32706)	