様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 6 月 13 日現在

機関番号：62615
研究種目：基盤研究(B)
研究期間：2008～2010
課題番号：20300059
研究課題名（和文）　最小ユーザフィードバックによるインタラクティブ情報収集

研究課題名（英文）　Interactive Information Gathering with Minimal User Feedback

研究代表者　山田 誠二（SEIJI YAMADA）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：50220380

研究成果の概要（和文）：人間と知的システムが協調しながら問題解決を行う知的インタラクティブシステムの実現のために，ユーザからのフィードバックを最小限に抑えてパフォーマンスを向上させる最少ユーザフィードバックの枠組みの提案し，その様々な要素技術を開発した．最小ユーザフィードバック実現のためには，少ないユーザフィードバックを最大限に利用する技術が必要であるが，より効率的な制約クラスタリングアルゴリズム，類似度判定に適した GUI (Graphical User Interface)の基礎調査研究，人間の能動学習を促進する GUI，独立成分分析による非階層的クラスタリングの初期値決定法などを開発し，その有効性を実験的に検証した．
研究成果の概要（英文）：We proposed a framework of Minimal User Feedback that makes traditional interactive systems more intelligent. Also we developed various elementary techniques: an efficient algorithm of constrained clustering, GUI to judge similarity between two documents, to facilitate human's active learning, and a seeding method of k-means with Independent Component Analysis. Finally we provided experimental evaluation for them.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	6,300,000	1,890,000	8,190,000
2009 年度	4,300,000	1,290,000	5,590,000
2010 年度	4,300,000	1,290,000	5,590,000
総　計	14,900,000	4,470,000	19,370,000

研究分野：人工知能，知的インタラクティブシステム，インタラクティブデータマイニング，機械学習
科研費の分科・細目：情報学・知能情報学
キーワード：知的インタラクティブシステム，最小ユーザフィードバック，制約クラスタリング，人工知能，可視化

１．研究開始当初の背景

これまで人工知能，計算知能の分野でさまざまな知的システムが研究，開発されてきたが，それらのほとんどは自律的な知的システムの構築を目指しており，人間は最初に一度入力を与えるだけで，あとはシステムが人間に代わって問題解決を行うことを目標としている．しかし，そのような自律的システムには限界がある．知的システムが一度の処理でユーザを満足させる結果を出すことは困難だからである．

このような認識から，人間（ユーザ）と知的システムが協調して問題解決を行う知的インタラクティブシステム IIS (Intelligent Interactive Systems) の構築が，今後の計算知能，人工知能，特にデータマイニング，機械学習の実用化において現実的な解決となると考えられる．

しかしながら，IIS 実現のためにもいくつかの重要な課題がある．そのうち最重要なも

のの 1 つが，「インタラクションにおける人間（ユーザ）の負荷が高すぎる」という問題である．IIS ではユーザがシステムの出力を評価するが，その評価自体がユーザの負担になってしまう．

２．研究の目的

以上の背景から本研究課題では，IIS 構築手法として最少ユーザフィードバックの枠組みを提案し，それを実現する様々な要素技術の開発，評価を行うことを研究目的とした．

最小ユーザフィードバック MUF(Minimal User Feedback)とは，ユーザからのフィードバックを最小限に抑えて，かつパフォーマンスを従来以上に維持する枠組みであり，前述の IIS 構築のための重要課題を解決し，IIS 設計の指針を与える．

そして，最小ユーザフィードバック実装に必要な要素技術として，効率的な制約クラスタリングアルゴリズム，類似度判定に特化した GUI(Graphical User Interface)開発のための調査研究，人間の能動学習を促進する GUI，独立成分分析による非階層的クラスタリングの効果的シーディング技術を開発し，その有効性を実験的に検証する．

３．研究の方法

本節では，最小ユーザフィードバックの枠組み，またその実現のために必要な個々の要素技術について方法論を説明する．

(1) インタラクティブデータマイニング・情報収集／検索

IIS は，図 1 に示すインタラクティブデータマイニング・情報検索システムをベースに，それをより知的にすることを目指している．図 1 では，最初にユーザがデータマイニング・情報収集／検索システムに，初期入力（クエリ，制約）を与え，それを基にシステムがマイニング／検索結果をユーザに提示する．そして，その結果をユーザが自身の選好によって評価（ユーザフィードバック）する．次にそのユーザフィードバックを制約，訓練データとして再度処理を行い，その結果をユーザに提示，ユーザはユーザフィードバックを返すというループを，ユーザが満足するまで繰り返す．

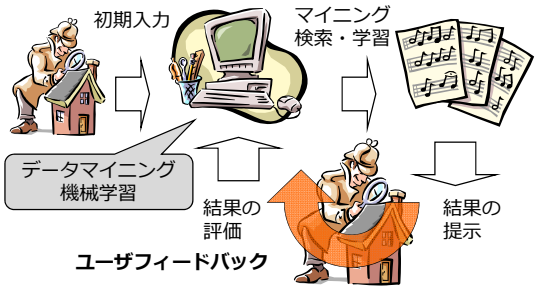

図 1 インタラクティブ
データマイニング・情報収集システム

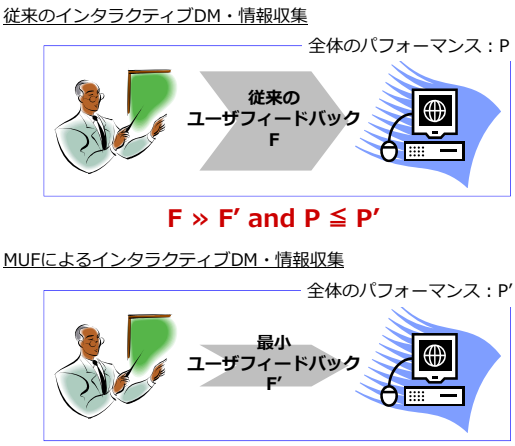

図 2 最小ユーザフィードバック

(2) 最小ユーザフィードバックの枠組み

IIS 実現の重要課題は、ユーザフィードバック自体がユーザに大きな負荷となることである．例えば，対話的文書検索でのユーザフィードバックでは，20～40 文書を精読して関連文書の判定を行うが，このユーザフィードバックは多大な認知的負荷となる．

よって，システム全体のパフォーマンスを落とさずに，このユーザフィードバックを小さく抑える技術の開発が，IIS 実現の必須条件と考える．この IIS が満たすべきこの条件を「最小ユーザフィードバック MUF (Minimal User Feedback)」と呼ぶ．図 2 に MUF のコンセプトを示す．そこでは，知的インタラクティブシステム全体のパフォーマンスを P, P’で表し，ユーザフィードバックを F, F’で表している．従来のシステムこのように，システム全体のパフォーマンスを維持しつつ（図中の P ≦ P’），ユーザフィードバックをできる限り抑える（図中の F ≫ F’）ことを目指す．

下記のようにユーザフィードバックを計算論的コスト，認知的コストの両側面から捉え，それぞれのコストを軽減する方針を採る．

- ユーザフィードバックの計算論的コストの最小化：少ないユーザフィードバックでも従来法以上のパフォーマンスを出す学習・データマイニングアルゴリズムの開発する．
- ユーザフィードバックの認知的コストの最小化：ユーザフィードバックの認知的負荷を抑える GUI を開発する．

以上 2 つの最小化により，ユーザフィードバックの量を抑え，かつ個々のフィードバックの認知的付加を抑えることができ，全体として MUF が実現される．

４．研究成果

ここでは，本研究課題において，MUF 実現

$$
\begin{aligned}
\min_K: \quad & \bar{L} \bullet K \\
s.t.: \quad & k_{ii} = 1, && i = 1,\ldots,n \\
& k_{ij} = 1, && \forall (i,j) \in M, \\
& k_{ij} = 0, && \forall (i,j) \in C, \\
& k_{ir_t^j} \le -\bar{l}_{ir_t^j},\ k_{jr_t^i} \le -\bar{l}_{jr_t^i}, && \forall (i,j) \in M \\
& k_{ir_t^j} \ge -\bar{l}_{ir_t^j},\ k_{jr_t^i} \ge -\bar{l}_{jr_t^i}, && \forall (i,j) \in C \\
& K \succeq 0
\end{aligned}
$$

図 3 類似度の最適化問題

のために我々の開発した具体的な要素技術について，その方法論と研究成果を説明する．

なお，これらの要素技術は，データマイニングや情報検索で最も広く利用されている（制約）クラスタリングを対象タスクとしている．

(1) 制約伝播による制約付きクラスタリングの改良

通常のペアワイズ制約を近傍のデータにも伝播させることで，少数制約でも高い性能をもつ制約付きクラスタリングを開発した．制約付きクラスタリングとは，通常のクラスタリングに，「同じクラスタに入って欲しいデータペア」，「同じクラスに入って欲しくないデータペア」などの制約をユーザが与え，それらの制約をできるだけ満たすようにクラスタリングを行うことである．制約の表現は，一般的であるペアワイズ制約である Must リンク（同じクラスタに属すべきデータペア），Cannot リンク（同じクラスタに属してはいけないデータペア）を扱い，制約付きクラスタリングアルゴリズムとしては，グラフラプラシアンと半正定値計画問題を利用した距離学習をベースに用いる．

距離学習によるクラスタリングでは，できるだけペアワイズ制約を満たすように，図 3 の丸枠以外のような定式化を行う．ここでは，K が求める距離行列，Lがグラフラプラシアン，M は Must リンク集合，C が Cannot リンク集合である．この最適化によって制約をできるだけ満たすような距離行列が計算される．これに対して，図 4 のように，あるデータペア D1, D2 の Cannot リンク，あるいは D2, D3 間の Must リンクがあった場合に，D2 の k 最近傍のデータ（図中 $k = 2$ で N1 と N2）にもそのペアワイズ制約を伝播させることで，少ない制約を拡張することが可能となる．図 3 における丸枠の部分が拡張された制約である．これらは，Must リンクは近傍データも同様に近づく，Cannot リンクは近傍データも同様に遠ざける制約となっている．このような制約の拡張により，少ない制約でも精度の高い制約付きクラスタリングを実現できる．

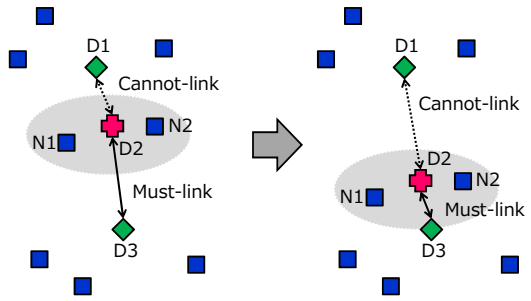

図 4 ペアワイズ制約の伝播

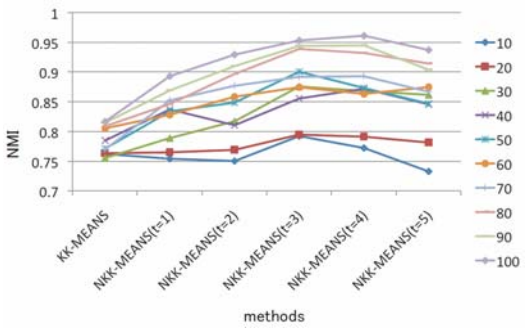

図 5 比較実験の結果

また，従来法との様々な比較実験によって，その有効を検証した．図 5 に UCI レポジトリでの比較を示す．ここでは，制約数が 10～100，KK-MEANS が従来手法，NKK- MEANS(t = n) が n 近傍まで伝播させた提案手法で，y-座標の評価は正規化相互情報量(NMI)を用いている．

(2) 認知的付加の低い制約判定の GUI

制約クラスタリングのユーザフィードバックユーザは，ユーザが 2 つの対象の類似性を判定する必要がある．その認知的負荷を下げる GUI 設計のために，文書の類似性判定において，判定者に提供する手がかりによる判定コストの違いについて調査研究を行った．手がかりとして，元文書，単語，スニペットを用いた．本実験では，文書対が与えられたときに，下記に示す 2 種類の単語・スニペットを提示している．

・単語 : TFIDF の高い単語のうち，両文書に共通して含まれる単語（common term），一方の文書のみに含まれる単語（specific term）
・スニペット : common term を多く含むスニペット（common snippet），specific term を多く含むスニペット（specific snippet）

大学院生 18 名，学部生 2 名，研究者 1 名を含む計 21 名に実験に参加してもらい，各手がかりを用いて 3 回ずつ，計 9 文書対の類似性判定を行ってもらった．

慣れを排除するため，各手がかりを初めて用いた場合，3 回目に用いた場合それぞれに分けて分析を行った．判定時間に関して多重比較検定を行った結果，1 回目の手がかり利用時では元記事を用いた場合が単語，スニペットを用いた場合よりも有意に長く，3 回目

表 1 類似性判定実験結果（正解率）

手がかり	元記事		スニペット		単語	
利用時	1st	3rd	1st	3rd	1st	3rd
正解(人)	12	17	11	16	10	11
不正解(人)	9	4	10	5	11	10
合計(人)	21	21	21	21	21	21

の利用時では元記事を用いた場合が単語よりも有意に長いという結果が得られた．

正解率に関する実験結果を表 1 に示す．利用時は，1 回目（1st）あるいは 3 回目（3rd）の手がかり利用時の別を表している．カイ二乗検定を行った結果では，1,3 回目とも優位な差はなかったが，1 回目と 3 回目の実験結果を比較すると，元記事，スニペットにおいて正解率が向上する傾向がみられた．

手がかりにより正解率の向上効果が異なる理由について考察するために，被験者の手がかり閲覧行動を視線追跡装置で分析した．単語およびスニペットを閲覧した時の典型的な視線移動を図 6 に示す．図において，二つの文書の手がかりは左右に並べて配置されている．また，上段に common term / common snippet，下段に specific term / snippet が提示されている．図より，単語閲覧時は二つの文書に関する手がかりを交互に閲覧していることがわかる．これは，単語の場合は個々の単語を独立に吟味できるのに対し，スニペットや文書の場合は単語列がコンテクストとして意味を持つという違いがあるためと考える．単語の場合は高速に吟味可能である反面，文書対によっては判定を行うのに情報が不足する場合があり，手がかりになれた後（3 回目）でも正しく判定することが困難であったと考える．

また，「同じトピックの文書を対で提示した場合は common term / snippet を，異なるトピックの文書を提示した場合は specific term / snippet を多く閲覧する」傾向も見られた．すなわち，被験者は同一トピックに属するかどうかの仮説を立てた後，その仮説を検証するのに必要な情報を重点的に吟味することが示唆される．

図 6 単語閲覧時（左），スニペット閲覧時（右）の視線の移動

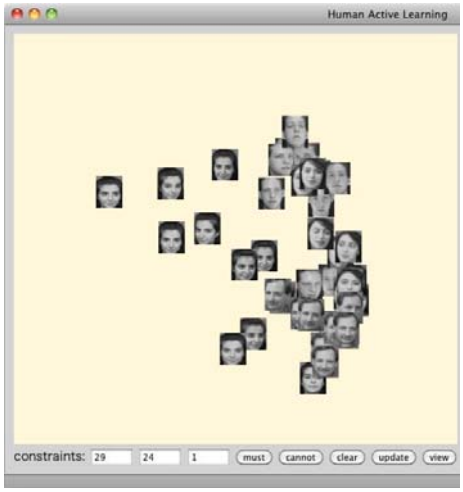

図 7 ユーザの能動学習のための GUI

(3) 人の能動学習を促進する GUI

ユーザに判定してもらう未判定データをいかに選択するかに関する研究である能動学習は，機械学習，特に分類学習において期待情報利得をベースとして研究されている．一方，近年人間の能動学習の能力にも注目が集まりつつある．

最小ユーザフィードバックの「ユーザフィードバックの認知的コストの最小化」を目指して，特にユーザの能動学習を促進する GUI の研究を行った．制約付きクラスタリングを対象とし，ベースとなるクラスタリングシステムはオンライン距離学習を使って，制約が与えられる毎に制約付きクラスタリングを行う．ペアワイズ制約は，図 7 のような GUI を通してユーザから与えられる毎にクラスタリングが更新される．

図 7 の 2 次元インタフェースにデータを表示するために，多次元尺度法の固有ベクトルのペアを用いるが，様々な観点を提示してユーザの能動学習を促進する目的で，複数の固有ベクトルのペアを x 軸, y 軸として用意し，提示する．この 2 軸の変更はマウスクリックにより簡単にできるので，ユーザはいろいろな観点から次にどの制約を与えればよいクラスタリングが可能かを考えることを促す効果が期待できる．

実際に，この GUI によるインタラクションを通じて，「大きなクラスタ中で離れているデータペアの Must リンクを選択」という能動学習のヒューリスティックスがユーザによって獲得され，それがランダム選択，uncertainty sampling（分類学習における能動学習のヒューリスティックス）よりも有効であることを実験的に示した．

(4) 独立成分分析による k-means の初期値決定

非階層的クラスタリングの代表的アルゴ

リズムである k-means 法は, 初期値である最初のクラスタ中心に依存して結果が大きく変わるという欠点をもつ．この問題を克服するため，最小ユーザフィードバック実現の要素技術として，高速な非階層的クラスタリング手法が安定してよい解を得ることのできる初期値の決定法を独立成分分析(ICA)により開発した．

クラスタリングと ICA との対比では, クラスタリングのデータ集合を ICA の観測信号と考え, ICA によって求められる復元信号 Y がクラスタ同士を独立させるようなクラスタの特徴ベクトルであると考える．つまり, この手法では, 与えられたデータ X から ICA によって，各クラスタを特徴づける独立成分ベクトル IC を求め，IC からコサイン距離が最も近いデータ点を初期値として設定する．以下にアルゴリズムを示す.

a. データ X から，ICA を用いて k 個の独立成分 IC_t, $t \in \{1,\cdots, k\}$ を得る.
b. 全ての独立成分 IC_t に対して，次式の F_{ICA} を最小にするデータ $\mathbf{x}_j$ をクラスタ中心 $\mathbf{c}_i$ として選択する.

$$F_{ICA} = \frac{\mathrm{IC}_t \bullet \mathbf{x}_j}{\left\|\mathrm{IC}_t\right\|\left\|\mathbf{x}_j\right\|}$$

c. クラスタ中心を k 個選択した後，k-means 法と同様の処理を行う.

この方法を用いることで，クラスタ同士の独立性が高い k-means 法のクラスタ中心の初期値を設定できる.

この手法の有効性を検証するために，UCI レポジトリと大規模なデータセットを用いて実験を行った．実験では小規模データとして，UCI ベンチマークデータと Open Directory Project コーパスを用いて，k-means 法の初期値決定の従来法である KKZ 法, K-means++法, そして主成分分析に基づく方法との実験的比較を行った．評価方法は，正規化相互情報量であった．その結果，提案方法は，従来法と同等あるいはそれらを凌駕する性能を示した.

以上，4 つの要素技術を総括するに，これらの技術を組み合わせることにより，最小ユーザフィードバック実現に必須の条件である「ユーザフィードバックの計算論的コストの最小化」と「ユーザフィードバックの認知的コストの最小化」が満たされ，最小ユーザフィードバックによる知的インタラクティブ情報収集・データマイニングが実現されることとなる.

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 9 件）

1. M. Okabe and S. Yamada: An Interactive Tool for Human Active Learning in Constrained Clustering, *Journal of Emerging Technologies in Web Intelligence*, Vol.3, No.1, pp. 20-27 2011.（査読あり）
2. M. Chen, S. Yamada and Y. Takama: Investigating User Behavior in Document Similarity Judgment for Interactive Clustering-based Search Engines, *Journal of Emerging Technologies in Web Intelligence*, Vol.3, No.1, pp. 3-10, 2011.（査読あり）
3. P. Li and S. Yamada: Extraction of Web Site Evaluation Criteria and Automatic Evaluation, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 14, No. 4, pp. 396-401, 2010.（査読あり）
4. M. Okabe and S. Yamada: Learning Similarity Matrix from Constraints of Relational Neighbors, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol.14, No.4, pp. 402-407, 2010.（査読あり, *JACIII Young Researcher Award* を受賞）
5. 高間康史, 山田隆志 : 時空間的動向情報の探索的分析を支援するインタラクティブな情報可視化システム, 人工知能学会論文誌, Vol. 25, No. 1, pp. 58-67, 2010.（査読あり）
6. 高間康史, 瀬尾優太 : 可視化表現共有型掲示板システムによる地域防犯活動議論支援, 知能と情報, Vol. 21, No. 3, pp. 316-326, 2009.（査読あり）
7. 松井正一, 山田誠二 : 遺伝的アルゴリズムによる階層メニューの最適化, 人工知能学会論文誌, Vol. 23, No. 6, pp. 494-504, 2008.（査読あり）
8. T. Onoda, H. Murata and S. Yamada: SVM-based Interactive document Retrieval with Active Learning, *New Generation Computing*, Vol. 26, No.1, pp. 49-61, 2008.（査読あり）
9. Y. Takama, A. Matsumura, T. Kajinami: Interactive Visualization of News Distribution in Blog Space, *New Generation Computing*, Vol. 26, No. 1, pp. 23-38, 2008.（査読あり）

〔学会発表〕（計 14 件）

1. M. Okabe and S. Yamada: Constrained Clustering with Interactive Similarity Learning, In Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems (SCIS&ISIS-2010), pp. 1295-1300, Okayama, Japan, 11 Dec. 2010.
2. T. Onoda, M. Sakai and S. Yamada: Seeding

Method Based on Independent Component Analysis for k-Means Clustering, In Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems (SCIS&ISIS-2010), pp. 1306-1309, Okayama, Japan, 11 Dec. 2010.
3. Y. Takama, C. Minghuang and S. Yamada: Document Similarity Judgment for Interactive Document Clustering, In Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems (SCIS&ISIS-2010), pp. 1310-1315, Okayama, Japan, 11 Dec. 2010.
4. H. Murata, T. Onoda and S. Yamada: A Kernel for Interactive Document Retrieval Based on Support Vector Machines, In Proceedings of Joint 5th International Conference on Soft Computing and Intelligent Systems and 11th International Symposium on Advanced Intelligent Systems (SCIS&ISIS-2010), pp. 1316-1321, Okayama, Japan, 11 Dec. 2010.
5. S. Yamada and K. Kobayashi: REBO: A Life-Like Universal Remote Control, In Proceedings of World Automation Congress (WAC-2010), IFMIP-200, 6 pages, Kobe, Japan, 20 Sep. 2010.
6. Y. Takama, M. Chen and S. Yamada: Effect of Snippet on User's Relevance Judgment of Documents, In Proceedings of World Automation Congress (WAC-2010), IFMIP-170, 6 pages, Kobe, Japan, 20 Sep. 2010.
7. M. Okabe and S. Yamada: An Interactive Tool for Constrained Clustering, In Proceedings of World Automation Congress (WAC-2010), IFMIP-562, 6 pages, Kobe, Japan, 20 Sep. 2010.
8. Y. Takama: Snippet Generation for Document Similarity Judgment, International Seminar on Intelligent Systems, 8 pages, Kosice, Slovakia, 5 Sept. 2010.
9. M. Okabe and S. Yamada: An Interactive Tool for Constrained Clustering with Human Sampling, In Proceedings of the International Workshop on Intelligent Web Interaction 2010 (IWI-2010), pp. 108-111, Toronto, Canada, 31 Aug. 2010.
10. M. Chen, S. Yamada and Y. Takama: Analysis of User Feedback Cost for Document Similarity Judgment, In Proceedings of the International Workshop on Intelligent Web Interaction 2010 (IWI-2010), pp. 87-90, Toronto, Canada, 31 Aug. 2010.
11. T. Onoda, M. Sakai and S. Yamada: Careful Seeding based on Independent Component Analysis for k-means Clustering, In Proceedings of the International Workshop on Intelligent Web Interaction 2010 (IWI-2010), pp. 112-115, Toronto, Canada, 31 Aug. 2010.
12. M. Okabe and S. Yamada: Clustering with Constrained Similarity Learning, In Proceedings of International Workshop on Intelligent Web Interaction 2009, pp. 30-33, Milano, Italy, 15 Sept. 2009.
13. M. Okabe and S. Yamada: Interactive Spam Filtering with Active Learning and Feature Selection，Proceedings of the International Workshop on Intelligent Web Interaction 2008，pp. 165-168, Sydney, Australia, 9 Dec. 2008.
14. M. Okabe and S. Yamada: Spam filtering with Active Feature Identification，Proceedings of 4th International Conference on Soft Computing and Intelligent Systems，pp. 1218-1223, Nagoya, Japan, 20 Sept. 2008.

〔図書〕（計 1 件）

1. 原島（監修），山口，久保田，高間：インテリジェントネットワークシステム入門，コロナ社，2008.

６．研究組織

(1) 研究代表者

山田 誠二（YAMADA SEIJI）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：50220380

(2) 研究分担者

小野田 崇（ONODA TAKASHI）
(財) 電力中央研究所・システム技術研究所・上席研究員
研究者番号：40371661
高間 康史（TAKAMA YASUFUMI）
首都大学東京・システムデザイン研究科・准教授
研究者番号：20313364
岡部 正幸（OKABE MASAYUKI）
豊橋技術科学大学・情報メディア基盤センター・助教
研究者番号：50362330