様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 2 月 28 日現在

機関番号：12601
研究種目：特定領域研究
研究期間：2006～2010
課題番号：18049013
研究課題名（和文）情報爆発時代におけるサイバー空間情報定量評価基盤の構築

研究課題名（英文）Building Quantitative Evaluation Platform for Exploding Cyber Information

研究代表者
喜連川　優（KITSUREGAWA MASARU）
東京大学・生産技術研究所・教授
研究者番号：40161509

研究成果の概要（和文）：本研究は、情報源の中でも最も増加率の高いウェブ情報源に対して各種解析手法の有効性を定量的かつ再現性を持たせた形で評価する定量的評価基盤を構築することを目的とする。超高速クローラ（ウェブページ収集システム）を改良し、細粒度時系列差分スナップショットを取得し各種テキストインデクス、種々のリンクインデクス等を実装した高度プラットフォームを構築した。その上で再現性のある定量的評価手法を実現するために様々なサイバー空間の解析を行った。

研究成果の概要（英文）：In this research, we have built a quantitative evaluation platform for comparing various analysis techniques for the Web that growth is highest among other information sources. We improved our own high performance crawler, and collected massive snapshots of the Web over time. Then, text and link indices have built on the snapshots, and various utilities were provided. On this platform, we performed various analyses for realizing reproducible evaluations.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	16,200,000	0	16,200,000
2007 年度	10,500,000	0	10,500,000
2008 年度	10,800,000	0	10,800,000
2009 年度	15,400,000	0	15,400,000
2010 年度	13,600,000	0	13,600,000
総　計	66,500,000	0	66,500,000

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：情報システム、コンテンツ・アーカイブ、 計算機システム、データストレージ、情報工学、 ウェブ情報、リンク解析、性能評価

１．研究開始当初の背景

近年人類の創生する情報は爆発的に増加しており、膨大な情報源から真に必要とする情報を如何に抽出するかという課題の解決は急務となっている。近年の知識労働者はその仕事の時間の 30%以上を必要な情報を「探す」という行為だけに費やしていると報告されている。我国が今後国際競争力を維持する為には知的集約度の高い技術を創出し続ける必要があり、膨大な情報源から必要なものを探すという行為の抜本的効率化が不可欠と言える。

２．研究の目的

本研究では情報爆発時代における情報検索の新たな基盤技術として、情報源の中でも最も増加率の高いウェブ情報源に対して定量的評価基盤を構築することを目的とする。

即ち、サイバー空間からの情報獲得に関して種々の研究が過去なされてきたものの、ウェブでは刻々とコンテンツが変化することから、例えば、現行のサーチエンジンと比べより良い結果が得られていることを再現性のある形で定量的に示すことは不可能であった。本研究では、各種手法の有効性を定量的かつ再現性を持たせた形で評価するプラットフォームを構築する。具体的には年限内に、既に独自に開発した超高速クローラ（ページ収集システム）を改良し、複数面の大規模スナップショット、時系列差分スナップショットを取得し各種テキストインデクス、特徴的グラフ様態に着目した種々のリンクインデクス、並びに各種ユーティリティを実装した高度プラットフォームを構築する。

３．研究の方法

サイバー空間情報定量評価基盤の設計及び実装を行い、その上で再現性のある定量的評価手法を実現するために、不完全なウェブスナップショットにおける定量的評価尺度の提案、ウェブスパムの挙動解析等、様々なサイバー空間の解析を行った。また、大規模データ解析を支える基盤ソフトウェアを構築した。

４．研究成果

（1）不完全な大規模ウェブスナップショットにおけるページ新規度の推定

大規模かつ変化の激しいウェブの観測においては、コンテンツの新規性、寿命、信頼性などを決定する際に、様々な不確定要素がつきまとう。まず、すべてのウェブページをリアルタイムに収集しつくすことは本質的に不可能であるため、各スナップショットの適切なサイズを決定することができない。米 IBM の調査では、10 億ページを収集した後でも、未収集のページ数が収集済みのページ数を上回ることが示されている。これは、各スナップショットの収集を何時止めるべきか判断できないことを示している。全ページの収集が不可能なことから新規に収集したページが本当に新しいとは限らないという不確定要素も生じる。新しく収集したページがそれ以前に収集されていなかった場合でも、前回の収集においてクローラがなんらかの理由でそのページに到達できなかった可能性があるため、確実に新しいとは言い切れない。また、同様の理由により、収集できなかったページがウェブから消滅したかどうかを判断することも難しい。最新のクロールにおいて収集されなかったページは何らかの理由で到達できなかった可能性があり、アクセスができなかったページであっても一時的にサーバが停止していた可能性が残るからである。ウェブの進化を観測する既存の研究においては、収集するページの集合や、サイトの集合を固定して定期的な収集を行うことで定量的な評価を実現しているが、これらの手法は収集範囲を厳しく制限するため、無視できない量の新規出現情報を見落とす上、現実のウェブ空間の進化とはかけ離れた変化を追跡することになる。

本研究では、このように本質的に不完全な性質を持つ大規模時系列差分スナップショットにおいても定量的な評価を可能とするようなフレームワークの構築を目指して、ウェブページの新規性に関する不確定性を解消する新規度という概念を提案した。新規度は時系列差分スナップショットにおいてある時点のページがどの程度の確信を持って新規に出現したかを判断するための尺度であり、時系列的に変化するウェブのリンク構造を解析することで算出することができる。新規度を利用することで、新規に出現したページを高い精度で抽出することが可能になる。

各スナップショットが収集された時期を、$tk(1 < k < n)$ とし、$W(tk)$ を時間 tk に収集されたページの集合とする。各時間のウェブスナップショット $W(tk)$ においてウェブページの集合 $V(tk)$ とハイパーリンクの集合 $E(tk)$ からなるグラフを $G(tk) = (V(tk),E(tk))$ と表す。$V(tk)$ は、$W(tk)$ 内のページからリンクされていれば、$W(tk)$ には存在しないページも内包する。これら $W(tk)$ 外のページは少なくともそのページが存在していたことを示す（図 1）。

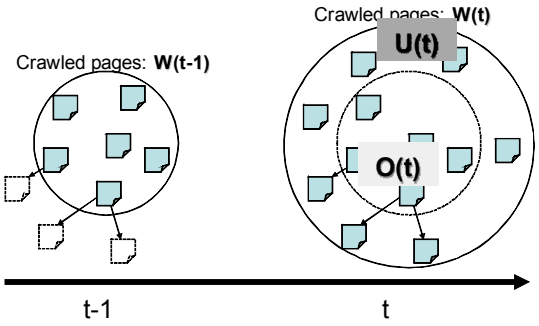


図 1：ウェブスナップショットの時間差分

新規度の算出のためにまず、$W(tk)$ 内のページを古いページ $O(tk)$ および新規度が不明なページ $U(tk)$ に分類する。ただし、$W(tk) = O(tk) \cup U(tk)$ and $O(tk) \cap U(tk) = \varnothing$ となるようにする。$O(tk)$ は、tk-1 以前に存在していたページ群を指し、$U(tk)$ は、それ以前のスナップショットに存在せず、tk に新しく収集されたページを指す。

あるページ p の新規度 $N(p)$ は、$U(tk)$ 内のページについて定義され、p がどの程度の確信度で、tk-1 と tk の間に新規出現したと判断できるかを示す。$N(p)$ は、0 から 1 の

間の値を取り、1 は、p が tk-1 と tk の間に新規出現したことを最も高い確信度で示し、0 は、p の新規度がまったく不明であることを示す。

新規度は、ページ p を指すリンクの新規度を用いて算出される。ここでまず、p は検索エンジンから取得可能になるまでウェブ上には存在しないと仮定する。すなわち、ページ p は、tk-1 と tk の間に新規出現したリンクのみによって指されている場合に、tk に出現したと仮定する。これは、p はクローラがリンクをたどって到達できるようになったときにウェブ上に現れたと考えることを意味する。

リンク(q, p) の新規度は、リンク元であるページ q の変化を調べることで判定することができる。

① リンク元ページ q が O(tk) に含まれ、tk-1 および tk 両時間に収集されているとき(q ∈ W(tk-1) かつ q ∈ W(tk))、リンク(q, p) は、tk-1 と tk の間に新規出現したと判定できる。(U(tk) の定義から、p は V(tj)(j < k) に含まれないため、q から p へのリンクが tk-1 になかったことが判る).

以降、過去 2 回(tk-1 および tk) にわたって取得されたページの集合を、L2(tk) ⊆ O(tk) と記す.

② q が O(tk) に含まれ上記の条件を満たさないとき、リンク(q, p) がいつ作られたか分からないため、その新規度は不確定である。このケースは、q が断続的に取得された場合、または q が W(tj)(j < k)の外側にあった場合に起こる。

③ q が U(tk) に含まれる場合、リンク(q, p) の新規度は、q の新規度に依存する。これは、q が新規である場合に(q, p) も新規であることを意味する。このリンクの新規度から、ページ p の新規度の再帰的な定義が以下のように導かれる。

$$\mathcal{N}(p) \quad (1 \quad \delta)\frac{\sum_{(q,p)\in I(p)} n(q,p)}{|I(p)|}$$

$$n(q,p) \quad \begin{cases} 1 & q \in I_2(t_k) \\ 0 & q \in O(t_k)\backslash I_2(t_k) \\ \mathcal{N}(q) & q \in U(t_k) \end{cases}$$

ただし、I(p) は p を指すリンクの集合を指す。パラメタ δ は、tk-1 以前に p へのリンクがスナップショット外において存在していた可能性を示し、新規度の伝播を減衰させる効果を持つ。この新規度の有効性を示すため、我国のウェブページを継続的に収集したアーカイブを用いて実験を行い、全てのページを新規ページとみなすベースラインよりも良い適合率および再現率で新規ページを抽出できることを確認した。

（2）サイバー空間におけるスパム挙動分析

近年、ウェブページのコンテンツやリンク構造を操作することにより、検索エンジンにおけるランキングを意図的に上昇させようとするウェブスパムという行為が顕著に見られるようになった。ウェブスパムには大きくわけて、タームスパム、リンクスパムと言う手法が存在する。タームスパムはウェブページに関係の無い単語を多数混入することで検索結果を操作する手法であり、リンクスパムはページ周辺に密なリンク構造を構築することで Google の PageRank のようなリンク解析に基づくランキングを操作する手法である。

ウェブスパムは、ウェブにおけるコンテンツおよびリンク構造に大きな変化をもたらしており、その挙動を明らかにすることは、サイバー空間情報の定量評価を行う上で非常に重要である。本研究では、特にリンク操作を用いたウェブスパムに着目し、580 万サイト、2 億 8 千万リンクからなる全日本のウェブサイトグラフに対して以下に示す 3 種類のグラフアルゴリズムを適用することで、スパム構造の分析を行った。これらの分析により、ウェブグラフ全体におけるリンクスパム構造の概要が明らかとなった。

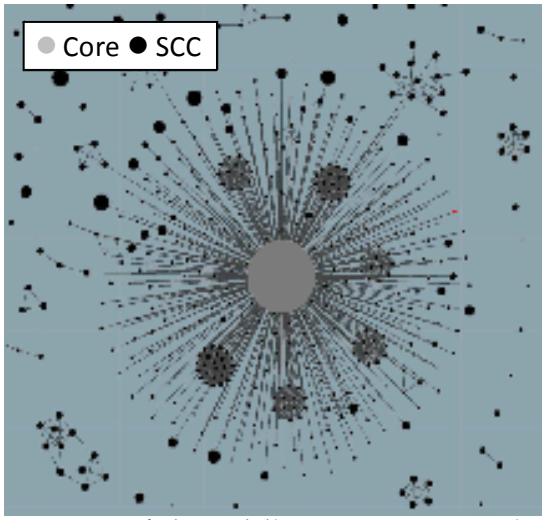


図 2：ウェブグラフ全体におけるスパム構造

①まず、サイトグラフを強連結成分(SCC: strongly connected component) に分解する。従来の研究から知られている通り、この分解によりコアと呼ばれる、全体の 30%程度のサイトを含む最大の SCC が発見された。さらに新たな発見として、コアの周辺に多数存在する大きな SCC が高い確度でリンクスパムであることが確認された。(図 2)

②コアの内部にはまだ多くのリンクスパムが含まれており、これを発見するためには強連結成分より密な構造を抽出しなければな

らない。ここでは極大クリーク列挙アルゴリズムをコアに対して適用し、抽出された極大クリークがやはり高い確度でリンクスパムであることを確認した。
③ここまでに発見した、SCC およびクリークは非常に強く結合された構造を持っており、その周辺には緩く結合されたスパムサイトがまだ多数存在する。最後に、SCC およびクリークをシンクとした最小カットを求めることで、周辺に存在するスパムサイトを抽出できることを確認した。

（3）サイバー空間におけるスパム・非スパム境界の検出手法

本研究では、特にリンク操作を用いたリンクスパムに着目し、スパムサイトと非スパムサイトの境界を検出する手法を提案する。リンクスパムの典型的な手法では、多数のサイトを用意しその間にリンクファームと呼ばれる非常に密なリンク構造を構築することで PageRank の向上を期待する。ただし、PageRank を最適化するためにはファームを構築するだけでなく、リンクハイジャックと呼ばれる手法でファーム外部のサイトから価値の高い被リンクを獲得することが必要である。リンクハイジャックには様々な手法が存在するが、もっとも単純な手法は、掲示板やブログのように自由にコメントを書き込めるページにスパムサイトへのリンクを投稿するものである。また、期限切れで開放されたドメインを購入することでそれまでにそのドメインに張られていたリンクを獲得する手法も存在する。こうしたリンクハイジャックがスパムサイトと非スパムサイトの境界となっており、これを検出することは、以下の点で重要である。
① ハイジャックされたサイトはそのまま放置すると続けて攻撃される傾向にある(コメントを繰り返し投稿される等)。検出したハイジャックサイトを監視することで新たに出現するスパムサイトを素早く知ることができる。
② サイトのランキングを行う際に、検出したハイジャックサイトからのリンクの重みを落とすことで、新たに出現するスパムサイトに対して頑健なランキングを行うことができる。
③ ウェブのクローリングを行う際に、検出したハイジャックサイトからのリンクをたどる優先順位を落とすことで、新たに出現するスパムサイトを収集することを避けることが可能になる。

本研究のハイジャックサイト検出手法は、各サイトについて、信頼性スコアおよびスパムスコアを算出し、スコアにリンクハイジャック特有の変化が起こるサイトをハイジャックサイトとして検出する。我々の実験では、既存の信頼性スコア、スパムスコアの中では、Gyongyi らの提案した Core-based PageRank を用いたものが最も境界を検出するのに適しているという結果が出ており、以降ではこのスコアを利用する。各サイト p について算出された、信頼性スコアを White(p)、スパムスコアを Spam(p) とする。通常のサイトでは信頼性スコアの方がスパムスコアより高くなっている(White(p) > Spam(p)) ことが期待され、スパムサイトではスパムスコアの方が高くなっている(White(p) < Spam(p)) ことが期待される。従って、これらのスコアの逆転が起こっているサイトでハイジャック行為が行われている可能性が高いと考えられる。

本手法では、対象とするサイトおよびリンク先サイトにおける信頼性スコア、スパムスコアの分布を考慮してハイジャックサイトを検出するためのハイジャックスコアを提案した。580 万サイト、2 億 8 千万リンクからなる全日本のウェブサイトグラフを用いて実験を行った結果、67.5%の適合率でハイジャックサイトを検出できることが分かった。さらに、抽出されたハイジャックサイトを調査することで、ハイジャック手法のうち、ブログ、掲示板、期限切れドメインを利用したものが半数を占めることが判明した。

（4）スパムリンク生成サイトの検出

本研究では、リンクハイジャック等の原因により、スパムへのリンクを生成し続けるようになったサイトをスパムリンク生成サイトと呼び、これを抽出する手法を提案した。具体的には、3 年間分の全日本ウェブアーカイブを用い、ある年のスナップショットにおいて、スパムへのリンクが増加したサイトを正解とし、それより前のスナップショットの情報のみを用いて、そのサイトを分類できるかという問題を解くことになる。スパムリンク生成サイトになるサイトを事前に知ることができれば、新たに出現するスパムサイトのサンプルを効率的に取得することが可能となり、新たなスパムに合わせたスパムフィルタの更新を素早く行うことが可能になる。

分類を行う際の、サイトの特徴量としては、入次数、出次数、PageRank、ハイジャックスコアなど様々な、リンクベースの特徴量に加え、その時間変化を考慮した。さらに、URL に含まれる部分文字列の頻度も考慮した。分類には、オンライン学習アルゴリズムである Average Passive-Aggressive アルゴリズムを用いた。

2004 年から 2006 年の 3 年分のアーカイブを用いた実験結果を図 3 に示す。全ての特徴量を用いた結果(ALL)で、97%を超える精度で分類ができており、これはリンクの特徴量のみ、又は URL の特徴量のみを用いた結果を

上回っている。

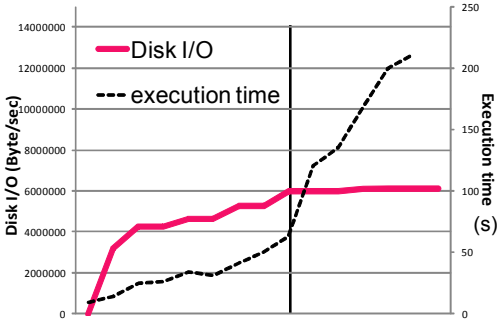
	2004				2005			
	P	R	F	AUC	P	R	F	AUC
ALL	**0.911**	**0.801**	**0.853**	**97.19%**	**0.857**	**0.848**	**0.852**	**97.22%**
Link-related	0.864	0.642	0.736	89.01%	0.442	0.704	0.543	88.89%
Link+Temp					0.653	0.573	0.611	92.67%
URL-related	0.948	0.699	0.805	90.04%	0.890	0.698	0.783	88.12%

図 3：スパムリンク生成サイト検出手法の評価

（5）データインテンシブアプリケーション実行時の負荷分散ミドルウェア

情報システムにおいて扱われるデータ量が爆発的に増加していることから、リソース使用状況に合わせてスケーラブルなシステム構築が可能なクラウドコンピューティングへの期待が高まっている。本研究においては大規模なウェブ情報処理等のデータインテンシブアプリケーションを対象とし、ローカルシステムの使用状況から判断して、リソースが不足している場合はスケーラブルに増減させた外部のクラウドリソースへ動的に負荷分散を行うミドルウェアを構築した。

図 4：実行時間と Disk I/O



全体の実行時間とクラウドの重量コストを最小にするためには、まずはローカルシステムを有効な範囲で使い切ることが必要となる。本研究で対象としているデータインテンシブアプリケーションでは通常の計算処理と異なり、CPU は I/O 待ちとなっていることが多い。そこで本研究では負荷の指標としてディスクアクセス量を用いる。

図 4 は実行するジョブ量を変化させた場合の、アプリケーションの実行時間と Disk I/O の値である。図 4 のように、実行時間はジョブ量が増えるに伴い長くなるが、Disk I/O はある一定の値を超えると飽和となり、ジョブが増えても値として増加しない状態に陥る。Disk I/O が飽和に達すると実行時間が極端に長くなる。そこで本ミドルウェアにおいてはこのように飽和した状態を「リソースを使い切った」状態とし、そのマシンへのジョブの投入を終了する。図 5 のようなシステム環境で本ミドルウェアの評価を行った。

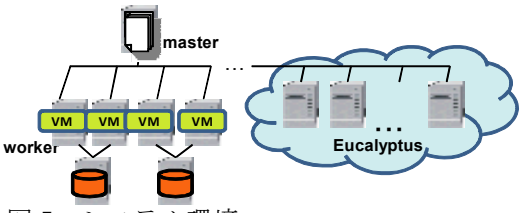


図 5：システム環境

クラウドにおけるストレージはローカルサイトのストレージを用い、クラウドから iSCSI 遠隔ストレージアクセスを行う。図 6 にジョブを 50 回目まで投入した際の結果を示す。Disk I/O と対応して、ジョブの投げ分けが行われていることが分かる。遠隔 iSCSI を用いたにも関わらず、現実的な時間でアクセスが行われており、データ配置のフレキシビリティも考えると、クラウド利用の際の有望な実現方式である。

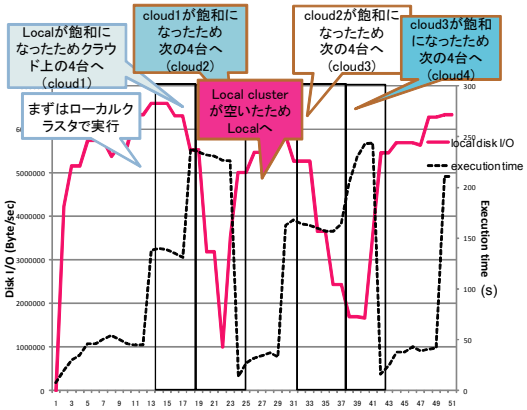


図 6：動的負荷分散の実験結果

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 19 件）

① Lin Li, Shingo Otsuka, Masaru Kitsuregawa: Finding Related Search Engine Queries by Web Community Based Query Enrichment. World Wide Web, Vol. 13, No. 1-2, pp. 121-142, 2010. 査読有

② Young-joo Chung, Masashi Toyoda, Masaru Kitsuregawa: Detecting Hijacked Sites by Web Spammer using Link-based Algorithms. IEICE Trans. on Information and Systems E93.D(6), pp. 1414-1421, 2010. 査読有

③ Yasuhiro Fujiwara, Yasushi Sakurai, Masaru Kitsuregawa: Fast Likelihood Search for Hidden Markov Models. ACM Transactions on Knowledge Discovery from Data(TKDD), Vol. 3, No. 4, pp. 18:1-37, 2009. 査読有

④ 豊田 正史: 検索エンジンスパムとその対策技術. 人工知能学会誌, Vol. 23, No. 6, pp. 760-766, 2008. 査読有
⑤ 相良 毅, 喜連川 優: 住所情報を用いた店舗名称のクリーニング手法. 電子情報通信学会論文誌 D, Vol. J91-D, No. 3, pp. 531-537, 2008. 査読有
⑥ 相良 毅, 喜連川 優: Web からの効率的な新規店舗の発見・登録支援手法. 情報処理学会論文誌:データベース, Vol. 48, No. SIG11, pp. 49-57, 2007. 査読有
⑦ 大塚 真吾, 喜連川 優: Web アクセスログとその利活用. 人工知能学会誌, Vol. 21, No. 4, pp. 410-415, 2006. 査読有

〔学会発表〕(計 105 件)
① Naoki Yoshinaga, Masaru Kitsuregawa: Polynomial to Linear: Efficient Classification with Conjunctive Features. The 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1542-1551, 2009.8.6. (Singapore) 査読有
② Young-joo Chung, Masashi Toyoda, Masaru Kitsuregawa: Detecting Link Hijacking by Web Spammers. The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 339-350, 2009.4.29. (Bangkok) 査読有
③ Lin Li, Zhenglu Yang, Ling Liu, Masaru Kitsuregawa: Query-URL Bipartite Based Approach to Personalized Query Recommendation. The 23rd National Conference on Artificial Intelligence(AAAI 2008), pp. 1189-1194, 2008.7.16. (Chicago) 査読有
④ Lin Li, Zhenglu Yang, Masaru Kitsuregawa: Using Ontology-Based User Preferences to Aggregate Rank Lists in Web Search. The 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008), pp. 923-931, 2008.5.21. (Osaka) 査読有
⑤ Zhenglu Yang, Lin Li, Botao Wang, Masaru Kitsuregawa: Towards Efficient Dominant Relationship Exploration of the Product Items on the Web. The 22nd National Conference on Artificial Intelligence (AAAI 2007), pp. 1482-1488, 2007.7.24 (Vancouver) 査読有
⑥ Masashi Toyoda, Masaru Kitsuregawa: What's Really New on the Web? Identifying New Pages from a Series of Unstable Web Snapshots. The 15th International World Wide Web Conference (WWW2006), pp. 233-241, 2006.5.24. (Edinburg) 査読有

〔その他〕
報道:
① 喜連川 優. 後編「ネットはいま」現実とネットの融合 (動画), asahi.com(朝日新聞社), 2008.11.28.
② 喜連川 優. ネットはいま 第 1 部 さがす ページの記憶, 朝日新聞 夕刊, 2008.11.10.
③ 喜連川 優, 豊田 正史. 「情報大爆発」どうさばく 過去からの変化を分析, 朝日新聞 be, 2008.7.5.

６. 研究組織
(1)研究代表者
喜連川 優 (KITSUREGAWA MASARU)
東京大学・生産技術研究所・教授
研究者番号:40161509

(2)研究分担者
小口 正人 (OGUCHI MASATO)
お茶の水女子大学・理学部・教授
研究者番号:60328036

中野 美由紀 (NAKANO MIYUKI)
東京大学・生産技術研究所・特任准教授
研究者番号:30227863
(H20→H22:連携研究者)

相良 毅 (SAGARA TAKESHI)
東京大学・生産技術研究所・助教
研究者番号:80302777
(H18→H19) (H20:連携研究者)

豊田 正史 (TOYODA MASASHI)
東京大学・生産技術研究所・准教授
研究者番号:60447349
(H19) (H20→H22:連携研究者)

(3)連携研究者
伊藤 正彦 (ITOH MASAHIKO)
東京大学・生産技術研究所・助教
研究者番号:60466422
(H21→H22)