

# 知識発見から知識共創へ

## －知識社会ダイナミズムの基盤技術－

大阪大学産業科学研究所 知能システム科学研究部門

鷲尾 隆

### 概 要

我々の住む知識社会では、あらゆる活動はユビキタスな情報ネットワーク基盤によって接続され、相互に影響し合いながら遂行されていく。これらの活動は、更に相互作用の中で新たな知識を共同創出し、それら知識の共有によって活動内容を自ら変革していく。これまでのデータマイニングをはじめとする情報処理技術は、情報通信ネットワークをフィールドとして、このダイナミズムを支える知識共創技術への脱皮を迫られている。本講演では、これまでの知識発見技術が知識共創技術へと発展しつつある姿を紹介し、更にそれらが知識社会の進化にもたらすインパクトの片鱗を展望する。

### 1. はじめに

インターネットをはじめとする情報通信ネットワークが社会に広く深く浸透する中で、我々の日常、ビジネス、その他あらゆる活動が、ネットワークを通じたコミュニケーションを直接、間接に利用し、その大きな影響を受けるようになっていく。殆どのネットワークは、人をはじめとする末端頂点間の瞬時の情報通信機能だけでなく、その情報の多くをサーバーなどのあちこちの頂点に蓄積し、一定以上の期間に亘り人々の間で共有する機能を持つ。近年は、この蓄積情報が量的に拡大しているだけでなく、ホームページ以外にもブログやSNS、掲示板、ウィキ、動画サイト、データベース、多くの商用サービスなどの多様性を増し、更に質的にも我々の活動にとって必須のものが増えている。

一方、我々人間は、昔から身の回りの流通情報や蓄積情報から、我々の活動、特にその中の認識、判断、意思決定に役に立つと思われるものを選択し、それらを利用しやすい形態に変換して知識として用いてきた。著名な経営学者であり未来学者でもあったピーター・ドラッカーが述べたように、そのような知識は、我々の日常、ビジネス、研究開発、世論形成を含むあらゆる活動の推進、改良、変革の源である[1]。このような個々の知識創出のプロセスは、個人や小集団、企業や政府のような階層組織、マスメディアのような一点集中システムに内包され、社会に新たな情報や知識が生産、流通、蓄積するというスパイラルが構成され、実社会進化のダイナミズムとなってきた。

最初に述べたように近年に至っては、ネットワーク社会の到来と Web2.0 と呼ばれるようなコンテンツ技術・情報共有技術の進歩[2]、更にはユビキタス社会[3]へ向けた進化によっ

て、人々は膨大な流通・蓄積情報を利用可能となってきている。しかも、情報検索やデータマイニング技術の発展によって、我々はそれら膨大な情報の中から効率的に自らの活動に役立つものを選択し、より価値の高い知識として利用できるようになった。更には、我々が得た情報や知識を広範に公開、発信することも可能となった。これによって、我々の情報の生産、流通、蓄積、及び知識の構成と利用が飛躍的に活発化したのみならず、新たに生産、構成された情報や知識も一層多くの人々に共有されるようになった。前述の個々の知識創出プロセスは、ネットワークを通じて広範囲の人々によって相互に影響しあいながら同時進行する「知識共創」へと組みこまれつつあり、実社会進化のダイナミズムもこの知識共創に深く依存するようになっている。

しかしながら、小集団や階層組織、マスメディアと同様に、ネットワークを通じた知識共創もその中で個々人の知識創出プロセスが個別連携されて実現している点では従来と変わりはない。この現状に対して、ネットワーク技術、特に情報検索やデータマイニングに、近年、地味ではあるが着実な変革が訪れつつある。それは、知識共創支援とでも呼ぶべき技術への転化である。情報検索やデータマイニングが、個々人レベルのデータからの知識発見や創出を支援、あるいは部分的に半自動化する技術であったのに対し、知識共創支援は、個々人の知識創出を束ね、人々全体としての知識創出を支援、あるいは部分的に半自動化する技術である。このような技術はまだまだ発展途上であり、今後の進展に期するところが大きい。以下では現状進められている研究の幾つかを紹介したい。その上で、最後に知識共創技術が知識社会の進化にもたらすインパクトの一部を展望したい。

## 2. 知識共創技術

### 2. 1 技術の多様性と可能性

知識共創を目指す技術の研究開発は漸く本格化した段階であり、知識共創技術という用語自体が学会などで確立しているわけではない。しかし、ネットワークを通じた知識共創の支援や半自動化という目的を考えるなら、潜在的には多様なアプローチが考えられる。この中で目立った研究動向は、ネットワーク上に蓄積されている情報コンテンツやそれを生みだした作者、組織の信憑性(credibility)や重要性(importance)、権威性(authority)、結節性(junction)、品質(quality)を、種々の情報を総合して評価する技術の開発である。このような技術は、現状では主に情報検索結果提示やネットワークセキュリティー、コンテンツ著作権保護の観点から研究されているが、膨大な人々が作成、公開した情報の中から信憑性、重要性、権威性、結節性、品質の高いものを選択し、価値の高い知識に変換して利用し合うことによって、知識共創プロセスの質と効率を大きく向上させることが期待される。

これらの技術分野に限っても、現状では情報検索やデータマイニング、認証技術、暗号化技術、情報通信関連社会制度など、様々なカテゴリに散在して地道に研究がなされてい

る。例えば、情報発信者や情報内容が信頼できることを電子認証や電子透かしなどによって外部から担保することや、同じくそれらの信用を高め情報の受け手を保護するための法律や社会的規制などの研究がなされている。しかし、これらは比較的大きな社会的責任を求められる人々や組織の情報発信が主な対象であり、日常の口コミや流行、マーケティング、ビジネス上の企画や研究開発アイデアの創出、不特定多数の人々から発せられる世論形成など、実社会進化のダイナミズムを生み出すすべての知識共創をカバーしきれない技術ではない。本格的な知識共創支援や半自動化のためには、社会的権威とは関係なく、ネットワークに蓄積された情報コンテンツやその発信者、組織の信憑性、重要性、権威性、結節性、品質を評価する技術の確立が不可欠である。

このような情報コンテンツや作者、組織を直接評価する技術の多くは、情報検索やデータマイニングの分野で研究が進められつつある。それらを技術的観点から大きく括るなら、2つに分けることができる。1つはテキストマイニングやパターン認識技術を拡張して、個別情報の特徴や相互の整合性から信憑性などを評価するコンテンツ信憑性分析技術である。2つめは、リンクマイニングなどに代表されるようにネットワーク上の情報相互の関係構造から、各情報の価値や質を評価して重要性や権威性、結節性、品質を評価する重要性伝播モデリング技術である。実際には、両者を適宜組み合わせることが必要となるが、以下ではこれら2つの類型に分けて代表的な技術を紹介する。

## 2. 2 コンテンツ信憑性分析技術

この技術は、更に情報の表層的特徴から信憑性(credibility)を評価する技術と、複数の情報間の不整合度から信憑性を評価する技術に分けることができる。前者は、情報コンテンツを構成する文章の文体や表現、語彙、ページタグなどの表層的特徴から、情報内容や発信者の専門性を含めた信憑性を評価するアプローチである。たとえば、誤字、脱字、文の係り受けなど、文章の表層的特徴を様々な角度で分析することで、書き手が文章の推敲にどの程度の労力を投入したかを推定できる。この推敲労力は、その文章を書くために費やした情報収集やそれに基づく思考の労力と高い相関があることが知られており、文章内容の信憑性を判断する上での大きな手掛かりとなる[4,5]。

また、情報コンテンツが対象とする分野の話題網羅度や、網羅している話題のメジャー度を計測することで、その情報の信憑性を評価するアプローチも研究されている[4]。網羅度は、評価対象とする情報コンテンツに含まれる文章中の語彙をキーワードとして、検索エンジンによってその関連分野の情報検索を行い、そこで検索された他のコンテンツの文章中の語彙との重なり度合によって計測可能である。また、メジャー度は、同じく検索された関連分野情報の多くに含まれる主要語彙との重なりや、検索エンジンで上位にランクされる関連分野の主要情報に含まれる語彙との重なりによって計測可能である。更には、感情(センチメント)を表す語彙の辞書を構築し、対象とする情報コンテンツ内の文章中の感情表現語彙に焦点を当て、情報発信者の話題に関する基本姿勢を評価する手法も研究

されている[6]。また、検索エンジン上位表示対策のためのタグ配置や隠しキーワードの存在なども情報発信者の姿勢と密接な関係があり、ページ構成に基づいてその信憑性を評価することも考えられている[7]。

これらに対して、複数の情報間の不整合度から各情報コンテンツの信憑性を評価する技術については、現状では語彙などの手掛かりを利用し難い画像や音声、映像情報の信憑性評価に関する研究が行われている[8]。まず、対象とする情報コンテンツに含まれる画像、音声、映像などを説明する周辺文章情報や関連情報から対象コンテンツ内容の推定を行う。次に、検索エンジンによって同等あるいは類似の文章情報やそれと関連づけられた画像、音声、映像を取得する。対象とする情報コンテンツと検索エンジンで取得された情報コンテンツを比較・分類して、メディア情報横断（クロスメディア）的に不整合度を計測する。他の取得情報との不整合度が小さければ、対象情報の信憑性は高いと評価できる。更に、画像や音声、映像に対して視聴者が付与した感想や評価コメント（視聴者アノテーション）を上記に加えて信憑性評価を行うこともできる。

これらコンテンツ信憑性分析技術の研究開発は、世界的にも我が国総務省の委託研究プロジェクト「電気通信サービスにおける情報信憑性検証技術に関する研究開発」[9]が中心となって推進しており、これが母体となって開催している **Workshop on Information Credibility on the Web (WICOW)** [10]は、主要な成果発表の場の1つである。更に、このプロジェクトの成果として、重要かつ信憑性の高いブログに、より容易により速く到達することを目的としたマニア／ファン指向ランキングを出力するブログサーチエンジンの実証実験サイトも公開されている[11]。また、上記ワークショップを併催した **ACM Conference on Information and Knowledge Management (ACM-CIKM)** [12]や **ACM Special Interest Group on Information Retrieval Home Page (ACM-SIGIR)** [13]などの国際会議には関連する研究発表が多い。

### 2. 3 重要性伝播モデリング技術

おそらく開発当時の研究者は明確に意図してはいなかったであろうが、重要性伝播モデリング技術に基づく知識共創技術の端緒となったのは、検索エンジンサービスとして世界で最も大きなシェアを握るようになった **Google** のホームページランキング技術である **Page Rank** アルゴリズムである[14]。これはホームページなど **Web** コンテンツ間のリンク構造から、人手を使わずに自動的に各ページの重要性(**importance**)を評価する技術である。その基本原理は、インターネット上のあるページが他のどのページからリンクされているかを調べ、重要な多くのページからリンクされているページは更に重要であると見なし、各ページの重要性の評価値 **Page Rank** によってページをランキングするものである。より具体的には、あるページの **Page Rank** をそのページにから発するリンク数で割った値が、そのページからリンクされた各ページの **Page Rank** に加算される。たとえば図1に示すように、左上の **Page Rank=100** であるページから2つのリンクが発している時、それから

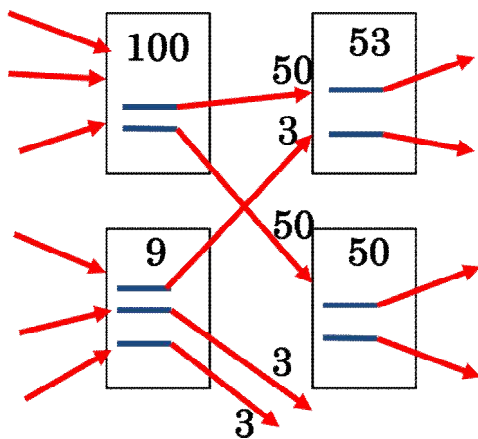


図1 単純化した Page Rank 計算例

リンクされている右上と右下の各ページの Page Rank に各々  $100/2=50$  が加算される。左下の3つのリンクが発している Page Rank =9 のページからリンクされている右上のページの Page Rank には、 $9/3=3$  が加算される。このようなページ間の重要性伝播を、対象とするすべてのページの Page Rank を値が収束するまで繰り返すことで、各ページの重要性を表す最終的 Page Rank を得ることができる。数値計算上は各ページの最終的 Page Rank は、ページ  $i$  への他のページからのリンク数を  $n_i$  とした時、ページ  $j$  からページ  $i$  へリンクがなされている場合に  $i,j$ -要素を  $1/n_i$ 、リン

クがなされていない場合に 0 とする確率遷移行列  $P$  の最大固有値に対応する固有ベクトルとなることが知られている。従って、対象とするネットワーク上のページ間のリンク関係から導かれる確率遷移行列の固有ベクトルによって、各ページの重要度は一意に決まる。Google は、この方法によってネットワーク上の数十億ものページを自動的にランキングし、ユーザーにとって重要性の高いページを優先的に検索結果に表示させることで、人々の知識創出プロセスを強力に支援する検索エンジンを提供した。これが Google の世界的成功の最大の要因である。

ネットワーク上のページを権威性(authority)や結節性(junction)の観点から評価する類似技術に、Hits(Hypertext Induced Topic Selection)アルゴリズムがある[15]。この技術では、情報の権威としての重要性の観点からページが他のハブページからリンクを張られている程度をオーソリティ値として評価し、また、情報の結節としての重要性の観点からページがオーソリティ値の高い他のページへ多くのリンクを張る程度をハブ値として評価する。より具体的には、ページにリンクしている他のページのハブ値の合計をオーソリティ値とし、ページがリンクしている他のページのオーソリティ値の合計をハブ値とする。例えば図2に示すように、1番目のページは他の3つのページからリンクを張られているのでオーソリティ値の初期値を 3 とするが、他のページへリンクしていないのでハブ値の初期値を 0 とする。同様に2番目のページは他のページからリンクを張られていないのでオーソリティ値の初期値を 0 とし、他のページへ4つのリンクを張っているのでハブ値の初期値を 4 とする。このようにして各ページのオーソリティ値とハブ値の初期値を求めた後、再度、各ページのオーソリティ値とハブ値を計算しなおす。これを各値が最終的な値に収束するまで反復計算する。数値計算上は各ページの最終的なオーソリティ値とハブ値は、ページ  $i$  からページ  $j$  へリンクが張られていれば  $i,j$ -要素を 1、張られていなければ  $i,j$ -要素を 0 と置いて得られる隣接行列を  $A$  とした時、それぞれ  $AA^T$  及び  $ATA$  の最大固有値に対応す

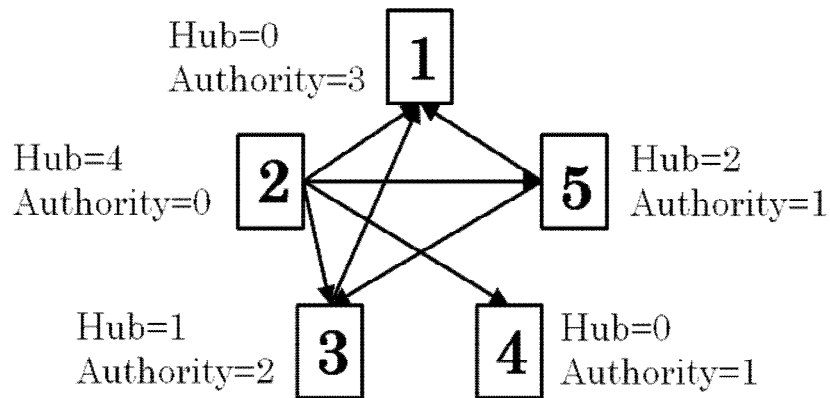


図2 オーソリティ値とハブ値の初期値計算例

る固有ベクトルの要素となることが知られており、各値はネットワーク上のページ間リンクの構造に関して一意に決まる。このアルゴリズムも権威性や結節性の観点から、検索エンジンなどの検索結果の自動的ランキングに用いることができる。

これら Page Rank アルゴリズムや Hits アルゴリズム自体は、単にネットワーク上の情報の重要性を種々の観点から評価する技術に過ぎない。しかし、これによって検索エンジンの上位に示される情報は多くの人々の目に触れて吟味・改良され、その内容が優れていれば更に多くからリンクを張られ益々重要性を増していく。このような自己組織化的なメカニズムによって、広範囲の人々が相互に影響しあいながら同時進行する知識共創を行うことが可能となる。

ネットワーク上の情報の関係に着目する重要性伝播モデリング技術に関する研究は、近年、更に多様な発展を遂げつつあり、その適用範囲も広がりを見せつつある。その1つにネットワーク上のオープン百科事典である Wikipedia を対象にした、原稿の品質(quality)と著作者や査読者の権威性(authority)を評価する手法の研究が挙げられる[16]。この基本原理は、前述の Hits アルゴリズムと似ている。即ち、原稿のクオリティ値はその著作者や査読者のオーソリティ値の総和であり、著作者や査読者のオーソリティ値は彼らが著した、または査読した原稿のクオリティ値の総和であると考えられる。ただし、Hits アルゴリズムはページ同士のリンク関係のみを対象とするのに対して、Wikipedia の場合には原稿のみではなく、著作者や査読者という主体も陽に扱わねばならない。しかも、1つの原稿を複数の著作者や査読者が部分的に著したり査読したりすることが多く、状況がやや複雑である。そこでまず、Wikipedia に現れる原稿  $i$  のクオリティ値を  $Q_i$ 、その原稿  $i$  に現れる個々の語彙  $k$  のクオリティ値を  $q_{ik}$ 、著作者または査読者  $j$  のオーソリティ値を  $A_j$  とする。そして、各原稿  $i$  のクオリティ値は、その原稿内の語彙のクオリティ値の総和、すなわち

$$Q_i = \sum_{k \in i} q_{ik}$$

と見なす。更に、語彙のクオリティ値  $q_{ik}$  はその語彙を原稿  $i$  に書き入れた著作者または査

読した査読者  $j$  のオーソリティ値の総和である、すなわち

$$q_{ik} = \sum_{q_{ik} \leftarrow j} A_j$$

とする。ただし、 $q_{ik} \leftarrow j$  は著作者または査読者  $j$  が語彙  $k$  を原稿  $i$  の中に書き入れた、または査読したことを表す。更に著作者及び査読者のオーソリティ値は、彼らが著した、または査読した原稿のクォリティ値の総和、すなわち

$$A_j = \sum_i c_{ij} Q_i$$

とする。ここで、 $c_{ij}$  は原稿  $i$  の中で  $j$  が書き入れた、ないしは査読した単語の数である。これら 3 つの式から  $A_j$  を代入消去すると

$$Q_i = \sum_{k \in i} \sum_{q_{ik} \leftarrow j} \sum_i c_{ij} Q_i \Rightarrow Q_i = M_Q Q_i$$

となり、 $Q_i$  を代入消去すると

$$A_j = \sum_i c_{ij} \sum_{k \in i} \sum_{q_{ik} \leftarrow j} A_j \Rightarrow A_j = M_A A_j$$

となる。ただし、 $M_Q, M_A$  は行列である。この問題の場合にも、前述の各アルゴリズムと類似して、 $M_Q, M_A$  の最大固有値に対応する固有ベクトルから、それぞれクォリティ値  $Q_i$  及びオーソリティ値  $A_j$  が得られる。

このように重要性伝播モデリング技術の根本原理は、多くのネットワーク上の情報やその発信者を評価する様々な指標の計算に用いることができ、今後、適用範囲をますます広げていくことが予想される。この技術に関連する研究発表は、前述の ACM-CIKM [12] や ACM-SIGIR [13] に加えて、International Conference on Data Engineering (ICDE) [17], SIAM Conference on Data Mining (SDM) [18], ACM International Conference on Web Intelligence (WI) [19], ACM International Workshop on Web Information and Data Management (WIDM) [20] など、多くの国際会議や国際ワークショップでなされている。

### 3. 知識共創技術のインパクト

ここで紹介した知識共創技術は、現在進められている種々の研究成果の一部に過ぎない。また、今後、その内容や適用範囲がどこまで拡大するかも、明確に予想することは困難である。しかしながらこの一連の技術は、従来の情報検索やデータマイニング技術と相まって、人々のネットワークを通じた広く多様な領域での知識創出の品質と効率を飛躍的に向上させる可能性を秘めており、我々の日常やビジネス、研究開発、世論形成などを含む今日の知識社会のあり方を多方面に亘って変えていくことはほぼ間違いないと思われる。この可能性の全貌をここで想像して議論することは困難であるため、最後に知識共創技術の発展によって予想し得る 1 つのインパクトの例を示すに留めたい。

現状、ネットワーク上で一般の人々が発する情報には、信憑性や重要性、品質の高いものと低いものが混在している。また、各情報発信者にも同様に信憑性や権威性が高い者と低い者が混在している。このため、流言飛語、誹謗中傷の類が横行し、現状ではプロ記者やディレクタによって運営される責任あるマスメディアが担うニュース報道や各種番組、記事の提供と同様なことを、一般の人々がネットワーク上で行うことは困難であると考えられている。一般人のブログを始めとするコンテンツ発信は、例えばその信憑性や重要性を自ら判別可能な人々で利用される、あるいは興味や知識を共有する限られたコミュニティーに向けた、いわゆるミニコミあるいはミニメディアとでもいうべきものが殆どである。しかし、ネットワーク上の情報の信憑性や重要性の評価や権威付けが自動的ないしは半自動的に実施可能で、それが情報検索エンジンやデータマイニングシステムと組み合わせられた形でポータルサイトなどに提供されれば状況は一変する可能性がある。一点集中型、トップダウン型のマスメディアに対し、一般の人々の情報からボトムアップに構成されるニュース報道や各種番組、記事などを提供するミドルメディアが世論形成に大きな影響を持つようになる可能性がある[21]。

現状の知識共創技術はまだ未熟であり、社会的な障壁もあって、このようなミドルメディアが容易に表れるとは言えない。しかし、パーソナルコンピュータやデジタルカメラ、Google などの検索サービスの例に見られるように、破壊的技術の性能は当初は既存の製品やサービスには及ばないものの、やがては肩を並べる存在にまで成長し、市場や社会を作り変えてしまう。情報検索、データマイニング、知識共創の技術が、このような大きな変化の引き金を引く資格は十分にあるように思われる。

#### 参考情報

- [1] Peter F. Drucker, 明日を支配するもの—21世紀のマネジメント革命, 上田 惇生 (翻訳), 1999.
- [2] Tim O'Reilly, What Is Web 2.0?, <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>, 2005.
- [3] 総務省, 平成15年情報通信白書, p.25, 2003.
- [4] Yutaka Kidawara, Information Credibility Analysis of Web Content, Second Workshop on Information Credibility on the Web (WICOW 2008), 2008.
- [5] Irit Askira Gelman and Anthony Barletta, A "Quick and Dirty" Website Data Quality Indicator, Second Workshop on Information Credibility on the Web (WICOW 2008), 2008.
- [6] Yukiko Kawai, et al., Using a Sentiment Map for Visualizing Credibility of News Sites on the Web, Second Workshop on Information Credibility on the Web (WICOW 2008), 2008.

- [7] Reyn Nakamoto, Reasonable Tag-Based Collaborative Filtering, Second Workshop on Information Credibility on the Web (WICOW 2008), 2008.
- [8] Satoshi Nakamura, et al., Can Social Annotation Support Users in Evaluating the Trustworthiness of Video Clips?, Second Workshop on Information Credibility on the Web (WICOW 2008), 2008.
- [9] 総務省平成 19 年度 電気通信サービスにおける情報信憑性検証技術に関する研究開発, [http://www2.nict.go.jp/q/q265/s802/info/20071026koubo/theme\\_b002\\_koubo.pdf](http://www2.nict.go.jp/q/q265/s802/info/20071026koubo/theme_b002_koubo.pdf)
- [10] Second Workshop on Information Credibility on the Web (WICOW 2008), <http://www.dl.kuis.kyoto-u.ac.jp/wicow2/>
- [11] マニア／ファン指向ランキングを有するブログ検索エンジン  
<http://kizasi.jp/labo/fansearch/index.py>
- [12] ACM 17<sup>th</sup> Conference on Information and Knowledge Management (ACM-CIKM 2008), <http://www.cikm2008.org/index.php>
- [13] ACM Special Interest Group on Information Retrieval Home Page (ACM-SIGIR), <http://www.sigir.org/>, The 31st Annual International ACM SIGIR Conference, <http://www.sigir2008.org/>
- [14] Lawrence Page, et al., The PageRank Citation Ranking: Bringing Order to the Web, Stanford University, SIDL-WP-1999-0120, 1999.
- [15] Jon Kleinberg, Authoritative sources in a hyperlinked environment, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [16] Meiqun Hu, et al., Measuring Qualities of Articles in Wikipedia, ACM Sixteenth Conference on Information and Knowledge Management (CIKM 2007), 2007.
- [17] International Conference on Data Engineering (ICDE), <http://www.icde2008.org/>
- [18] SIAM Conference on Data Mining (SDM), <http://www.siam.org/meetings/sdm08/>
- [19] IEEE/WIC/ACM International Conference on Web Intelligence (WI)  
<http://datamining.it.uts.edu.au/conferences/wi08/>
- [20] ACM International Workshop on Web Information and Data Management (WIDM) <http://workshops.inf.ed.ac.uk/WIDM2007/>
- [21] 藤代裕之, イノベーションのジレンマに襲われるニュースメディア, NIKKEI NET, IT+PLUS: インターネット : 連載・コラム, 2008. <http://it.nikkei.co.jp/internet/column/gatoh.aspx?n=MMIT11000001082008>