Web上のテキスト情報の信頼性と有益性の評価システムに関する研究

研究代表者	中川　裕志	東京大学・情報基盤センター・教授
研究分担者	吉田　稔	東京大学・情報基盤センター・助教
	清田　陽司	東京大学・情報基盤センター・助教
連携研究者	二宮　崇	東京大学・情報基盤センター・講師
	宇津呂武仁	筑波大学大学院・准教授
研究協力者	佐藤　一誠	東京大学・大学院生
	野田　陽平	同上
	森　　竜也	東京電機大学・大学院生

１　研究の概要

研究目的　本申請テーマではWeb上の情報の信頼性と有益性の評価システムを開発する。ここで信頼性は、既存の権威ある知識と同じ用語で物事を説明していることと定義する。ただし、これだけでは、単純に既存の知識の繰り返しである可能性が高く、利用者にとって有益な知見を与えてくれることが保証できない。そこで、有益性を示すために有力と考えられる次の条件を加える。

(1)関連が密でない複数の知識を統合した見解を提示している

上記の既存の権威ある知識として Wikipedia を対象とすることにした。(1)の定義を実装するために、Wikipediaのカテゴリー構造を利用し、カテゴリーを要素するネットワーク構造から得られた特徴すなわち素性を利用して、対象とする記事の有益さを評価するシステムを構築する。また、Wikipediaは言語間の格差が大きいため、その信頼性は言語によって異なる。言語間に差異を評価するシステムを試作し、言語間格差の実情を調査する。

本年度の研究成果の概要　本年度は、下記の課題に取り組んだ。

- Webにおける人名検索結果の参照曖昧性解消: 口頭発表[3]
- Wikipediaを利用した情報ナビゲーション: 学会誌等[4]、口頭発表[5,6]、受賞[1]
- 統計的機械学習による文書分類、同義抽出など: 学会誌等[1,2,3]、口頭発表[1,4]、受賞[2]
- 多言語Webデータからの情報抽出: 口頭発表[1]

以下では、これらのうち、2番目のテーマについて詳しく説明する。

２　意外性のある知識発見のためのWikipedia カテゴリ間の関係分析

概要

本研究の目的は，Web上の情報の信頼性と有益性の評価であるが、評価が定まりつつあるWikipediaの情報を対象にすることによって当面の信頼性を確保する。問題はむしろ有益性である。専門家は自分の専門分野については熟知しているから、既に自身の専門分野内の有益な情報は知っていると考えるべきであろう。むしろ、自分の専門分野からかなり離れた分野における知識で自分の分野とつながりがある知識があれば、それらは有益と考えるであろう。専門家でない場合にも同じようなことが考えられ、概念的には遠く離れ、かつ関係が薄いとみなされていた2つの分野のかけ橋になるような知識が発見できれば、有益な知識を見出したと考えるだろう。言い換えれば、意外性のある知識ということになる。この研究では、上記のような考察により、Wikipediaにおいて意外性のある記事を探すことを目的とする。

英語版Wikipedia は2008 年8 月11日に250 万記事，日本語版Wikipedia は2008 年6 月25 日に50 万項目を超え，膨大な情報量を誇る百科事典として広く認知されている．日本語版Wikipedia は9 個の主要カテゴリの下に，サブカテゴリ，記事が関連付けられており，大規模なグラフ構造を成している．各項目はそれぞれ複数の親カテゴリを持っており，また，同義語はリダイレクトとして関係付けられている．Wikipedia の記事は，カテゴリシステムによってさまざまな観点からの分類がなされている．この特徴をうまく用いると，

0個別の記事からだけでは得られない意外な知識の発見につなげることができる．例えば，「オープンコーラ」という概念がある。「コーラ」は万人周知の飲料だが、その製造方法がオープンではない。一方、「オープンソース」という概念はソフトウェアの世界では良く知られたもので、コード内容が公開されている。この二つの概念、すなわち、飲料とソフトウェアという離れた概念のかけ橋として、製造方法が公開されているコーラという意味で「オープンコーラ」という概念があり、これは意外性が高いと考えられる。

以下では、このように直観的に規定した「意外性」という概念を、Wikipediaのカテゴリが持つネットワーク構造を利用して定義し、その定義に沿って意外な知識をWikipediaから発掘する手法とシステムについて述べる。

利用する特徴情報

我々は、Wikipedia のカテゴリネットワークの構造情報のうち、上記の直観を実現するために有効であると間がられる以下の情報（feature）に着目した。

・各項目が属するカテゴリの数(親カテゴリの数)
・各カテゴリが持つ子項目の数
・各項目に関する，共通の親を1 つ以上持つ項目の数(兄弟項目の数)
・各カテゴリセットが持つ子の数
・2つのカテゴリA,Bの間の距離、すなわちA,Bの共通親カテゴリを軽油した場合のA,B間の最短パス長

機械学習による識別

上記の特徴情報を用いて意外性のある情報を発見する手法として第一に検討すべきものは、意外な記事の集合とそうでない記事の集合という教師データを作成し、SVMのような教師あり学習によって識別器を学習する方法である。教師データとしては180記事の意外性のある記事と、そうでない記事180記事を人手で作成した。また、実際は意外性のない記事が圧倒的に多数なので、ランダムに記事を選択して意外でない記事とみなし、意外性なしの教師データの数を変えて実験してみた。SVMでこれらの教師データで学習し、4分割の考査検定を行った結果を次の表に示す。

表１（意外性があると人手判断したデータ）／（意外性があるとSVMが判定したデータ）

意外性のあるデータ（左）とないデータ（右）の数	180:180	180:1800	180:18000
線形カーネル	90.70%	4.17%	0%
2次多項式カーネル	90.70%	8.33%	0%
3次多項式カーネル	90.70%	12.50%	0%
ガウシアンカーネル	88.37%	4.17%	0%

この結果からみると、意外性のある記事が少数派（手作業の経験では1000分の1以下）の状況においてはSVMによる識別の方法は機能しないことが判明した。

回帰による順位付け

このように、意外性の有無でWikipediaの記事を分類することが判明したので、意外性の程度によって記事を順位付ける方法を検討する。統計的機械学習の観点からすれば、識別ではなく回帰によるモデルを考えることになる。意外性に関する正例と負例はfeatureを各次元に持つ多次元空間中に分布する。 そこで、これに意外性の有無を表す次元として有＝１、無＝０の値を持つ次元を加える。具体的にはn種類のfeatureで表現される各記事は次のようなベクトルで表現される。

意外性のあるデータ：$(1,\ \text{feature}_1,\ldots,\text{feature}_n)$
意外性のないデータ：$(0,\ \text{feature}_1,\ldots,\text{feature}_n)$

このデータを1次式で回帰すると以下のようなベクトル空間上でのイメージとなる。つまり、赤い円で囲まれた意外性にある記事と青い円で囲まれた意外性のない記事を上記にベクトルで表して線形回帰すると紫

の２本の直線で囲まれた回帰平面が得られる。図中に橙色の円で表現された新規の記事をこの紫色の回帰平面に射影し、その意外性を表す軸（図の縦方向の軸）の値を得れば、これがこの記事の意外性の度合いを表す数値を与える。

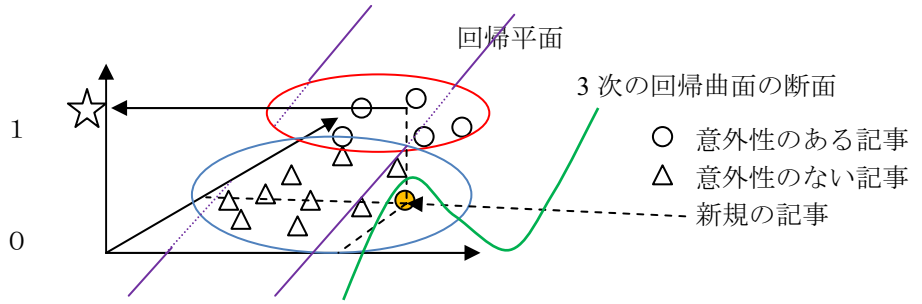

図１　意外性を表すベクトル空間における回帰

これは、1次の線形回帰だが、3次の曲面による回帰を用いれば緑色で断面を示すような曲面で回帰するため、意外性の有無の記事がより複雑な状況に散在しても対処できる可能性がある。このような方法で具体的にデータを処理した結果の例として、マイケルジャクソンの例を以下に示す。

例：カテゴリ：マイケルジャクソンに対して、それと遠い別のカテゴリおよび両方のカテゴリに共通する記事を調べてみた。意外性のある記事数とない記事数の比を１：１、１：１０、１：１００の3種類の場合で線形回帰および3次回帰した場合の上位3位までを下の表に示す。

線形回帰した場合

	順位	カテゴリＡ	カテゴリＢ	ＡＢ共通の記事
意外：非意外	1	マイケル・ジャクソン	エルヴィス・プレスリー	リサ・マリー・プレスリー
=1:1	2	マイケル・ジャクソン	ダンステクニック	ムーンウォーク
	3	マイケル・ジャクソンのアルバム	マイケル・ジャクソン	マイケル・ジャクソン
意外：非意外	1	福祉	マイケル・ジャクソン	USAフォー・アフリカ
=1:10	2	億万長者	マイケル・ジャクソン	ジェイ・レノ
	3	ダンス	マイケル・ジャクソン	ムーンウォーク
意外：非意外	1	福祉	マイケル・ジャクソン	USAフォー・アフリカ
=1:100	2	億万長者	マイケル・ジャクソン	ジェイ・レノ
	3	ダンス	マイケル・ジャクソン	ムーンウォーク

3次曲面回帰した場合

	順位	カテゴリＡ	カテゴリＢ	ＡＢ共通の記事
意外：非意外	1	福祉	マイケル・ジャクソン	USAフォー・アフリカ
=1:1	2	LGBTの人物	マイケル・ジャクソン	リサ・マリー・プレスリー
	3	ダンス	マイケル・ジャクソン	ムーンウォーク
意外：非意外	1	福祉	マイケル・ジャクソン	USAフォー・アフリカ
=1:10	2	LGBTの人物	マイケル・ジャクソン	リサ・マリー・プレスリー
	3	ダンス	マイケル・ジャクソン	ムーンウォーク
意外：非意外	1	福祉	マイケル・ジャクソン	USAフォー・アフリカ
=1:100	2	LGBTの人物	マイケル・ジャクソン	リサ・マリー・プレスリー
	3	ダンス	マイケル・ジャクソン	ムーンウォーク

これらの例で見られるように、回帰による方法は、意外性のある記事数とない記事数の比によらず安定した

結果を得られる。特に3次曲線での回帰は安定度が高い。これ意外の順位においても同様の安定度が得られた。

３　Wikipedia を用いた情報ナビゲータ

概要

Wikipedia のカテゴリ体系は、多様な観点を反映する集合知としての性質をもつ一方、図書館の分類体系などの学術用語体系との強いつながりをもっている。我々の研究グループではこの特徴を生かし、Wikipedia に含まれる一般的なキーワードを起点に、さまざまな観点での調べ方を提示し、信頼性の高い情報資源に誘導するシステムを構築している。本年度は、システムに入力されたキーワードと図書館分類体系との関連を示す**テーマグラフ**とよばれる図を自動的に描画する機能を追加した。また、本機能を国立国会図書館の Web サービス「リサーチ・ナビ」の検索コンポーネントとして実運用し、評価を行った。

情報の調べ方を推薦する要件

Web サーチエンジンの普及は、一般の人々の「情報探し」に対するものの見方に大きな影響を与えている。仕事や買い物、旅行、食事、読書、料理といった日々の行動をするにあたって「まずサーチエンジンで探す」という習慣は人々の間で広く共有されている。これにともない、「検索キーワードさえ適切に選べばサーチエンジンは答えを与えてくれる」とか、「サーチエンジンで何もみつからなければそれ以上探しても仕方ない」という誤解も生じかねない状況が生まれている。

しかし、私たちが日々抱く疑問の中には、検索キーワードを入力するだけでは答えを見いだせないものも数多く存在する。「関東大震災が発生したのはいつか？」「○○ってどんな病気？」という疑問に対しては、サーチエンジンで簡単に答えにたどりつくことができる。一方で、大学の学生の「関東大震災についてレポートを書かなきゃいけないんだけど、いったいどんな資料から調べたらいいの？」という疑問や、難病を抱えた患者の「自分の病気について最先端の治療を行っている病院を探す方法は？」という疑問に対しては、サーチエンジンは明解な答えは与えてくれない。

「レポートの書くための資料の探し方を知りたい」「病院を探す方法を知りたい」といった疑問を抱くのは、情報探しのテーマがあいまいな場合が多い。このような場合、情報探しのテーマを利用者に推薦することが求められる。情報探しのテーマを推薦するためには、サービス提供側は以下の 3 つの条件を満たしていることが必要だと考えられる（図 1）。

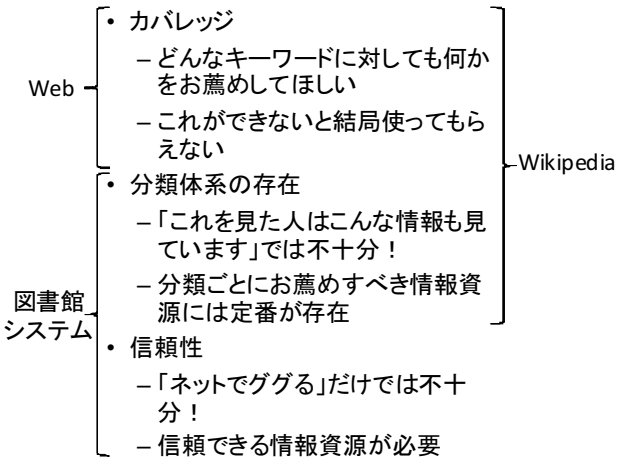

図 1　テーマ推薦の要件

1. カバレッジ
   あらゆる分野の質問に対して何らかの情報資源を推薦することが求められる。この条件を満たしていない限り、結局はサービスを利用してもらえない。
2. 分類体系の存在

お薦めすべき情報資源には、情報探しの分野に応じて「定番」とよべるものが存在する。例えば、図書館においては入門書(新書など)や参考図書がそれに該当する。「定番」となる情報資源を利用者に提示するためには、あらかじめそれらの情報資源を整理しておく分類体系が必要である。

3. 信頼性

どんな情報資源を推薦したとしても、利用者側が「このお薦めは信頼できる」という感情を抱かなければ結局は利用されない。「ネットでググる」ことで得られる情報資源を超える信頼性の根拠を利用者に示すことが重要となる。

日常の情報探しにおいて最も広く用いられている Web サーチエンジンは、上記の 3 つの条件のうち「カバレッジ」は満たしているが、「分類体系の存在」「信頼性」においては不十分である。1990 年代の Web では Yahoo!などのディレクトリサービスがこれらの条件を満たす役割を担っていたが、Web 空間の情報量が爆発している現代においては人手によるディレクトリの作成が追いついていない状況である。

一方、図書館の代表的な Web 情報サービスである OPAC では、「分類体系の存在」「信頼性」の条件は満たしているが、「カバレッジ」については不十分である。思いついたキーワードを OPAC に入力しても、何も情報が得られない場合はかなり多い。

オンライン百科事典 Wikipedia は、この 3 つの条件を考慮したとき興味深い位置に存在する。誰でも編集が可能であることから「信頼性」については課題があるが、「カバレッジ」「分類体系の存在」の条件は満たしていることから、Web と図書館システムの間のギャップを埋める架け橋として利用できる可能性がある。

Wikipedia と件名標目表を統合的に活用した分類自動導出

Wikipedia は「テーマ推薦の要件」と「カテゴリの構造」の観点からみたときに、きわめてユニークな特徴をもっている。この特徴をうまく用いて、情報探索の出発点として Wikipedia を利用し、そこから概念を一般化することによって図書館の分類体系に導いていくという方法を提案している。

図 2 に導出アルゴリズムの概要を示す。まず、Wikipedia カテゴリの構造について説明する。Wikipedia の記事「阪神・淡路大震災」には、カテゴリとして「日本の経済史」「地震の歴史」が付与されている。さらに、カテゴリ「日本の経済史」には上位カテゴリとして「経済史」が、カテゴリ「地震の歴史」には上位カテゴリとして「災害と防災の歴史」「地震」が付与されている。このように、Wikipedia の記事を一つとりあげてみると、関連するカテゴリ群をツリー構造として取り出せることがわかる。

次に、Wikipedia カテゴリと図書館の分類体系の対応付けについて説明する。Wikipedia カテゴリと図書館の件名の間には、カテゴリ名が一致するものが存在する。図 2 では、「経済史」「災害」「地震」が一致している。よって、「阪神・淡路大震災」につながるカテゴリが構成する有向グラフの構造を再帰アルゴリズムによってたどることで、「阪神・淡路大震災」に関連する分類を自動的に導出することができる。我々の研究グループでは、グラフのエッジに対する重みスコアをノード間の文字列類似度によって定義し、ビームサーチによって重みスコアが相対的に大きい件名を絞り込むアルゴリズムを採用している。

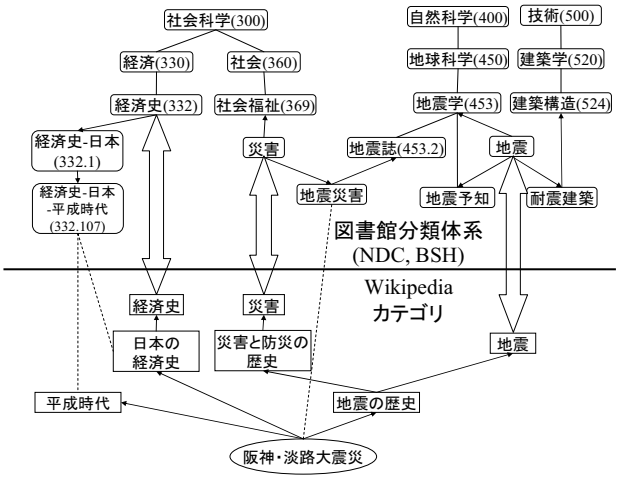

図 2 Wikipedia を利用した分類の自動導出

国立国会図書館リサーチ・ナビにおけるテーマグラフ生成

「リサーチ・ナビ」は、国立国会図書館が調査に役立つ情報を蓄積し、整理・体系化して Web 上に提供し、求める情報にユーザを案内するための Web サービスであり、2009 年 5 月より国立国会図書館の Web サイト http://rnavi.ndl.go.jp/ にて一般公開されている。さまざまな分野における有用な参考文献や情報源を簡潔にまとめた「調べ方案内」や、全国の図書館におけるレファレンス質問・回答事例を蓄積した「レファレンス協同データベース」、参考図書などの目次データベースなどのコンテンツを有している。リサーチ・ナビの検索画面例を図 3 に示す。

図 3　リサーチ・ナビの検索画面

リサーチ・ナビの検索システムには、入力されたキーワードからの分類自動導出の結果を「テーマグラフ」として自動的に描画する機能が実装されている。図 4 に、「燃料電池」という検索キーワードから導き出されたテーマグラフの例を示す。テーマグラフの内容を考察していくことで、「燃料電池」という概念がどのようなテーマと関連を持っているのかを知ることができる。例えば、「再生可能エネルギー」「循環型社会」「環境問題」などの件名と「燃料電池」との関連性を考察すると、「燃料電池は環境問題解決の切り札として注目されている」という背景を見いだすことが可能である。

評価

情報爆発サーチ共通ユーザ評価の一環として、被験者（50 名）にリサーチ・ナビ検索システムを利用してもらい、有用性についてのアンケートを実施した。その結果、50 名中 43 名（86%）の被験者は、テーマグラフから検索に有用な何らかのヒントを得たと回答した。

4　今後の展望

本年度は、まず Wikipedia から意外性のある情報を探し出すための支援手法について検討した。また、Wikipedia を情報ナビゲーションに利用する試みとして、Wikipedia で利用されている日常的な用語と専門文献の橋渡しを行う情報ナビゲーションシステムの構築と評価を行った。今後は、これらの研究を引き続き発

展させ、意外性があり信頼にたる情報を探し出す手法、システムの研究を目指す。

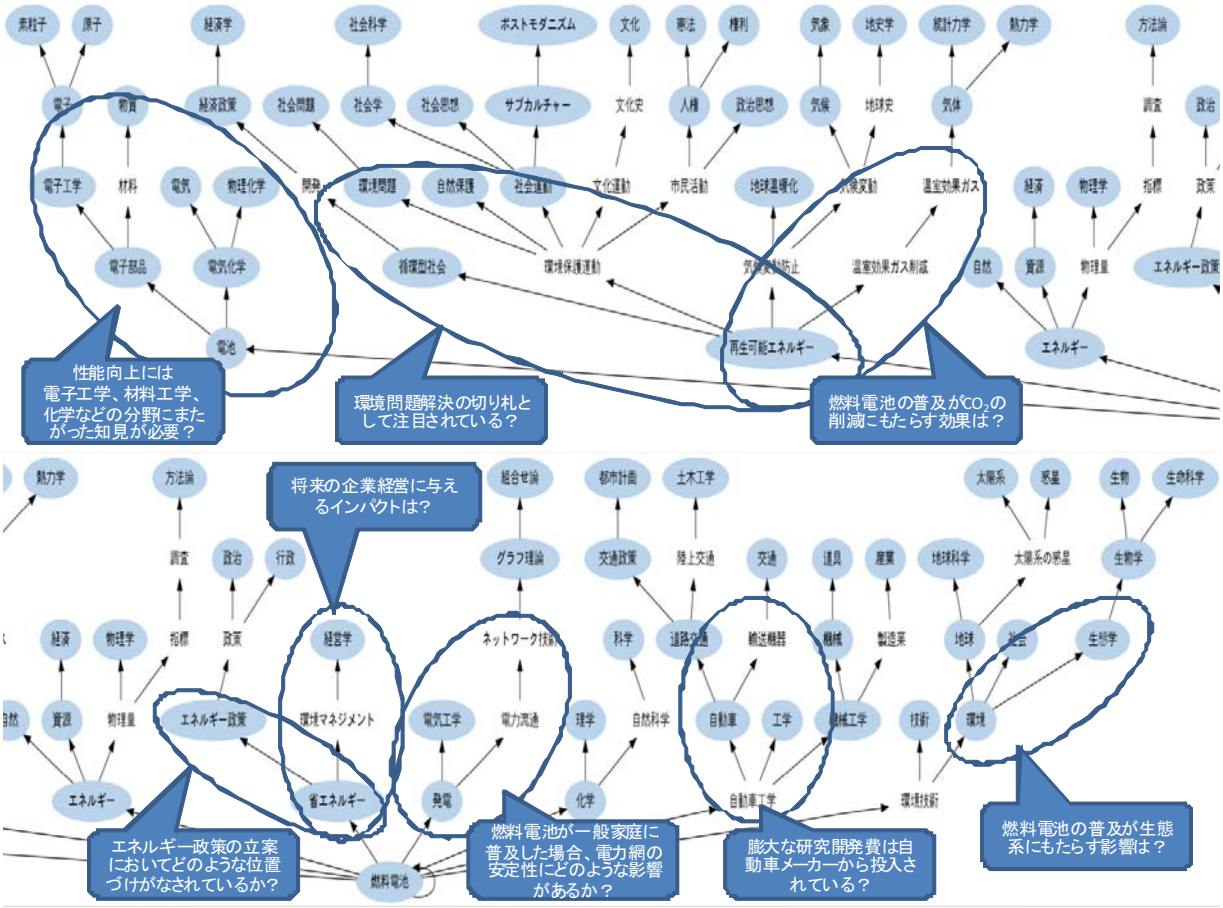

図 4　「燃料電池」のテーマグラフ

研究成果リスト

学会誌等

1. 吉田稔、中川裕志、寺田昭. “コーパス検索支援のための動的同義語候補抽出”, 人工知能学会論文誌　25(1), 2009（採録）
2. S.Matsushima, N.Shimizu, K.Yoshida, T.Ninomiya, H. Nakagawa. “Exact Passive-Aggressive Algorithm for Multiclass Classification Using Support Class”, SIAM SDM'2010(accepted).　April 29 - May 1, 2010.
3. Yo Ehara, Takashi Ninomiya, Nobuyuki Shimizu, Hiroshi Nakagawa. “Personalized Reading Support for Second-language Web documents by Collective Intelligence”, IUI2010, Hong Kong, China, Feb.7-10, 2010
4. Yoji Kiyota, Hiroshi Nakagawa, et. al.. “Exploitation of the Wikipedia Category System for Enhancing the Value of LCSH”. Joint Conference of Digital Library:JCDL 2009 , pp.411, 2009(poster)
5. Ninomiya, Takashi, Takuya Matsuzaki, Nobuyuki Shimizu and Hiroshi Nakagawa. (2009). “Deterministic shift-reduce parsing for unification-based grammars by using default unification”. In the Proc. of EACL 2009. pp. 603-611.
6. 清田 陽司. “リサーチ・ナビ検索システムの技術”, 参考書誌研究, Vol. 71, pp. 347-352, 2009.

口頭発表等

1. Hiroyuki Nakasaki, Yusuke Abe, Takehito Utsuro, Yasuhide Kawada, Tomohiro Fukuhara, Noriko Kando, Masaharu Yoshioka, Hiroshi Nakagawa, Yoji Kiyota. “Cross-Lingual Analysis of Concerns and Reports on Crimes in Blogs”, Mining User-Generated Content for Security (MINUCS2009), Venice Italy, Dec.9, 2009,
2. Issei Sato, Kenichi Kurihara, Shu Tanaka, Seiji Miyashita and Hiroshi Nakagawa .“Quantum Annealing for Variational Bayes Inference” The 25th Conference on Uncertainty in Artificial Intelligence (UAI2009) http://www.cs.mcgill.ca/~uai2009/proceedings.html , in Montreal, Canada, June 18-21 , 2009
3. Masaki Ikeda, Shingo Ono, Issei Sato, Minoru Yoshida and Hiroshi Nakagawa. “Person Name Disambiguation on the Web by TwoStage Clustering”. 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference, Madrid, Spain, April 21. 2009. (この競争型タスク（Web 人名検索）に参加した１７チーム中２位)
4. 佐藤一誠, 中川裕志. “Latent Dirichlet Allocation の量子アニーリング変分ベイズ学習”. IBIS2009. 2009
5. 森竜也, 増田英孝, 清田陽司, 中川裕志. (2009). Wikipedia エントリ構造抽出ツール: Wik-IE. 人工知能学会研究会資料 セマンティックウェブとオントロジー研究会 第 20 回 Wikipedia ワークショップ. 2009
6. 野田陽平, 清田陽司, 中川裕志. “意外性のある知識発見のための Wikipedia カテゴリ間の関係分析”. 人工知能学会研究会資料 セマンティックウェブとオントロジー研究会 第 20 回 Wikipedia ワークショップ. 2009
7. 中野 幹生, 緒方 淳, 清田 陽司, 東中 竜一郎, 翠 輝久: “【パネル討論】音声インタフェースにおける Web テキスト処理技術の利用”, 情報処理学会 研究報告 自然言語処理(SIG-NL) 191-12, 東京工業大学, 東京都, 2009.
8. 坂井 哲, 増田 英孝, 清田 陽司, 中川 裕志: 国立国会図書館リサーチ・ナビにおけるテーマグラフの生成, 情報処理学会 研究報告 情報学基礎(SIG-FI) 96-5, 秋葉原ダイビル, 東京都, 2009.
9. 清田 陽司: 図書館分類体系と Wikipedia を統合した情報探索支援システムの開発, 日本図書館研究会情報組織化研究グループ／情報知識学会関西部会 合同研究会, 大阪科学技術センター, 大阪市, 2009.
10. 清田 陽司: Wikipedia をいかに使いこなすか？ －知識抽出、情報ナビゲーション、そしてトピック発見－, 第 114 回 Ku-librarians 勉強会, 京都大学, 京都市, 2009.
11. 清田 陽司: Wikipedia と図書館情報資源のマッシュアップ, 第 36 回生物医学図書館員研究会, 順天堂大学, 東京都, 2009.
12. 清田 陽司: リサーチ・ナビ検索システムの技術, 第 11 回図書館総合展/学術情報オープンサミット 2009 フォーラム企画, 国立国会図書館主催, パシフィコ横浜, 横浜市, November, 2009.
13. 清田 陽司: 学生向けレファレンス支援ツールの可能性, 第 11 回図書館総合展/学術情報オープンサミット 2009 ミニ・フォーラム＆プレゼンテーション企画, 紀伊國屋書店主催, パシフィコ横浜, 横浜市, November, 2009.
14. 清田 陽司: 知識体系の新たな融合 : 情報探索と件名標目表の活用をめぐって, 京都大学図書館機構 平成 21 年度第 2 回講演会「次世代 OPAC を考える ―目録情報の視点から―」, 京都大学, 京都市, November, 2009.

受賞

1. 人工知能学会 2009 年度全国大会優秀賞:吉田稔, 中川裕志. Wikiwi: テキストマイニングによる Wikipedia 検索支援, 1C3-3, (6/17), 2009 年 10 月
2. 人工知能学会 研究会優秀賞: 佐藤一誠, 吉田稔, 中川裕志. 多重性を考慮したノンパラメトリックベイズグラフクラスタリングによる単語間関係抽出, SIG-DMSM-A801-07 (7/24), 2009 年 4 月