平成23年度採用分 特別研究員-DC 申請書

第 1 版

①申請資格	DC1	受付番号
②審査領域	工学(総合)	1016
③分　　科	情報学	
④細　　目	メディア情報学・データベース	⑥専門分野
⑤分科・細目コード	1004	ウェブ工学

筑波大学

1．申請資格等

（申請機関コード：0408）

（フリガナ） ⑦氏　　名	ヨシダ ミツオ 吉田 光男	⑧性　　別	男
⑨国　　籍	日本		
⑩生年月日	昭和59年9月27日　生（平成23年4月1日現在 26 歳）		

⑪学　　歴 （学部・修士）	1. 平成21年3月 筑波大学 第三学群 情報学類卒 2. 平成21年4月 筑波大学大学院博士前期課程 入学 （システム情報工学研究科 コンピュータサイエンス専攻）
⑫博士の状況	1.入学年月：　平成23年4月 入・進学見込 2.大学院名：　筑波大学(0408) 3.研究科名：　システム情報工学(099) 4.研究科種別：　研究科 5.専攻名：　コンピュータサイエンス専攻 6.課程種別：　博士課程（３年制） 7.休学期間合計：　0年 8.平成23年3月末時点における博士在学期間累計（休学期間を除く）：　0年 9.平成 23年4月時点における在学年次：　1年 10.博士に係る学歴の特記事項：　無 11.博士の追記事項
⑬研究・職歴等	1. 平成18年3月 ～ 平成23年3月(予定) 有限会社てっくてっく 代表取締役 2. 平成20年9月 ～ 平成21年3月 筑波大学産学リエゾン共同研究センター 客員研究員 3. 平成21年4月 ～ 平成22年3月 ティーチング・アシスタント(筑波大学) 4. 平成22年4月 ～ 平成23年3月(予定) ティーチング・アシスタント(筑波大学) 5. 平成22年4月 ～ 平成23年3月(予定) 筑波大学産学リエゾン共同研究センター 客員研究員

⑭日本学生支援機構等奨学金貸与の有無	有	⑮外国人留学生に対する奨学金等受給の有無	

⑯研究課題	時空間メタデータ検索をキーワード検索に統合したウェブ検索エンジンの実現

2. 現在までの研究状況 (図表を含めてもよいので、わかりやすく記述してください。様式の改変・追加は不可(以下同様))

① これまでの研究の背景、問題点、解決方策、研究目的、研究方法、特色と独創的な点について当該分野の重要文献を挙げて記述してください。

② 申請者のこれまでの研究経過及び得られた結果について、問題点を含め①で記載したことと関連づけて説明してください。
なお、これまでの研究結果を論文あるいは学会等で発表している場合には、申請者が担当した部分を明らかにして、それらの内容を記述してください。

ウェブ検索エンジンは、ウェブ上のコンテンツ急増に伴い、インターネットを利用する際には欠かせないサービスとなっている。インターネット技術の発展とともに、ウェブ検索エンジンは、会社のページなど一般的なウェブページのみならず、画像、動画、ニュース、ブログなど様々な情報を検索できることが求められている。一方、ウェブ上のコンテンツ急増は、ブログなど利用者が内容を生成するメディア（CGM）の普及に一因がある。我が国では、2004 年から 2005 年ごろにかけてブログコンテンツが急増しており、現在も増加傾向が見られる[1)]。しかし、十分な性能を達成できていないためか、ブログ検索の利用率が伸びていない[2)]。今後、ウェブコンテンツに占める一般的なウェブページの割合が大幅に低下すると予想でき、ウェブ検索エンジンは、ブログなどにあらわれる新しい種類のテキストコンテンツにも対応していく必要がある。

以上を踏まえ、申請者は、ウェブ検索エンジンの高度化に取り組んでいる。高度化のための主な着眼点は、コンテンツの自動抽出、時空間情報（時間情報・地理的位置情報）などのメタデータによる検索の 2 点である。以下では、これまでに取り組んできたコンテンツの自動抽出に関する研究について述べる。

右のブログページからもわかるとおり、ブログページは、ヘッダ、メニュー、広告、関連記事リストなど不要部分が多々存在しており、ページに占めるコンテンツ（ポスト・コメント）の割合が低い。そのため、ブログページのコンテンツを利用するためには、コンテンツの抽出が必要になる。

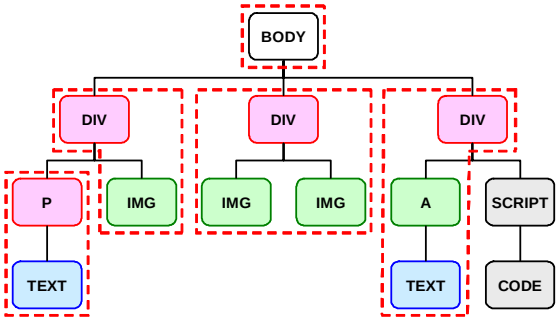
コンテンツ抽出手法の代表例として、人手によって抽出ルールを記述する方法が挙げられる。コンテンツの抽出を行う際は、ブログページごとにコンテンツの位置が異なることに留意する必要がある。同一のブログホスティングサービスを利用しているブログページに関していえば、コンテンツの位置が同一であると考えられるため、ブログホスティングサービスごとに抽出ルールを準備する必要があるものの、人手によって抽出ルールを記述する方法は有効に機能する。しかし、申請者による予備調査では、人気のあるブログサイトの少なくとも 30%がブログホスティングサービスを利用していないことが判明した。このような状況下では、人手によって抽出ルールを記述する方法には多大な労力を必要とすることは明らかであり、自動抽出手法を検討する必要がある。

以上の労力を軽減するために、事前に教師情報を準備しコンテンツ抽出のモデルを自動獲得する手法が提案されている。Bing ら[3)] は、平均的なウェブページにおいてどの位置にコンテンツが出現するかを学習する手法を提案している。これらの手法では、事前に教師情報を準備する必要があるため、その準備に労力を必要とする。また、平均的なウェブページの構造が変わると抽出が困難になるという問題を抱えている。

Lin ら[4)] や Debnath ら[5)] は、サイト内におけるページ中の部分の重要度スコアを計算することによりコンテンツを抽出する手法を提案している。これらの手法は、事前に教師情報を準備する必要がないものの、大きく分けて 2 つの問題点が存在する。1 つ目の問題点は、コンテンツ候補の抽出にコンテンツと不要部分を分断しやすい HTML タグのリストが必要であり、このリストはウェブページデザインの流行に左右されることである。2 つ目の問題点は、コンテンツを抽出するために重要度スコアの閾値が必要であり、各ページに適した閾値を人手で決定する事は、無数のウェブページが存在する今日では、非常に困難であるということである。

申請者らは、日本データベース学会論文誌において、コンテンツ候補（ブロック）の抽出に W3C が定義するブロックレベル要素を利用することにより、新たに HTML タグのリストを準備する必要のない手法を提案した。さらに、ある特定のウェブページにのみ出現するブロックはコンテンツであるという単純なアイデアを基に、各ウェブページに閾値を設定する必要がないコンテンツ抽出手法を提案した。（業績 1-1、全体を担当）

右の図は DOM ツリーから 5 つのブロックを抽出した例である。ブロックがコンテンツ及び不要部分の最小単位となるように、入れ子になっているブロックレベル要素を抜いて抽出する。ブロックがコンテンツであるか否かは、ウェブページ集合内でブロック同士を比較し、唯一出現するかどうかで判定する。提案手法をニュースサイトに対して適用すると、適合率 98.0%、再現率 91.1%という非常に高いコンテンツ抽出性能を示した。教師情報や閾値調整等の必要がないため、全自動でコンテンツを抽出することができる。



以降、上の提案手法（「従来手法」と呼ぶ）のさらなる改良について述べる。

申請者氏名　　吉田光男

DC

ブログのコンテンツは、ポストと呼ばれるブログの書き手によるコンテンツと、コメントと呼ばれるブログの読者によるコンテンツに大別する事ができる（図 1)。従来手法は、ポストとコメントを区別せずに抽出したが、ブログのコメントを利用する研究が行われ始めるなど、今後、ブログのコンテンツ抽出はポストとコメントを分離抽出することが期待される。

分離抽出する先行研究として、Cao ら[6)] は、ポストとコメントを分断する 1 本の境界線を発見する事により自動的に分離抽出する手法を提案している。しかし、コンテンツを 1 か所に特定するため、コンテンツの中に不要部分を多く含む傾向が高いという問題点がある。また、境界線の存在が前提であるため、コメントが付いていないブログ記事には適用が難しいという問題点もある。

筆者らは、Cao らの問題点を解決する手法として、情報処理学会研究報告（DBS）において、コンテンツのうち、ポストはブログ記事集合全てのブログ記事に出現するが、コメントはいずれかのブログ記事にしか出現しないというアイデアを基にした分離抽出手法を提案した（業績 4-4、全体を担当)。この提案の中で、要素識別子を基にしたコンテンツグループ化手法を提案した。さらに、言語処理学会第 16 回年次大会において、グループ化手法を応用して、コンテンツ抽出性能を向上させる手法を提案した（業績 4-11、全体を担当)。最終的な抽出性能は、ブログページに対して適用すると、コンテンツ 91.3%、ポスト 87.7%、コメント 87.4%という高い性能を達成できた（いずれも適合率と再現率の F 値)。

1) 総務省情報通信政策研究所. ブログ・SNS の経済効果に関する調査研究. 2009.
2) 総務省情報通信政策研究所. インターネット検索エンジンの現状と市場規模等に関する調査研究. 2009.
3) Lidong Bing, Yexin Wang, Yan Zhang, Hui Wang. Primary Content Extraction with Mountain Model. In Proceedings of IEEE CIT 2008, pp.479-484, 2008.
4) Shian-Hua Lin, Jan-Ming Ho. Discovering Informative Content Blocks from Web Documents. In Proceedings of ACM SIGKDD 2002, pp.588-593, 2002.
5) Sandip Debnath, Prasenjit Mitra, Nirmal Pal, C. Lee Giles. Automatic Identification of Informative Sections of Web Pages. IEEE Transactions on Knowledge and Data Engineering, vol.17, no.9, pp.1233-1246, 2005.
6) Donglin Cao, Xiangwen Liao, Hongbo Xu, Shuo Bai. Blog post and comment extraction using information quantity of web format. In Information Retrieval Technology: AIRS2008, pp.298–309, 2008.

3. これからの研究計画

(1) 研究の背景

2. で述べた研究状況を踏まえ、これからの研究計画の背景、問題点、解決すべき点、着想に至った経緯等について参考文献を挙げて記入してください。

申請者によるウェブ検索エンジンの高度化のための主な着眼点は、コンテンツの自動抽出、時空間情報（時間情報・地理的位置情報）などのメタデータによる検索の 2 点である。これまではコンテンツの自動抽出に関する研究を行ってきたが、これからはメタデータによる検索に関する研究を行う。なお、2 点は各々が独立するものではなく、コンテンツの抽出が正確に行えるほど、メタデータによる検索の性能が上がると予想している。

ウェブ検索エンジンの歴史において、革新的な出来事は Google の登場である。利用者の大半は検索結果の上位しか参照しないが[1)]、Google 以前のウェブ検索エンジンは TF-IDF[2)] を中心とする単純なアルゴリズムが支配的であった。Google は PageRank[3)] を導入することにより、性能を飛躍的に向上させることに成功した。しかしながら、ウェブ検索エンジン登場から 15 年が経とうとしている現在でも、あるキーワードが含まれるウェブページを探す事に重点が置かれたままである。現在のウェブ検索エンジンには、キーワード検索以外に、言語、地域、ファイルタイプ、日付などメタデータを指定する検索機能が準備されている。メタデータはウェブページの絞り込みに効果的であるものの、メタデータを指定する検索機能の利用率は高くない[4)]。

ブログなど利用者が内容を生成するメディアのテキストコンテンツには、利用者の行動記録が書かれる傾向があるが故に、時空間情報が含まれる場合が多い。コンテンツの充実に伴い、申請者は、ある時期（時間）や場所（地理的位置）に関するコンテンツを探したいという検索動機が伸びると予想している。

申請者は、例えば「ハチ公 ラーメン」で検索した時、渋谷駅周辺（「ハチ公」から推定、実際は数値情報）のメタデータを持つラーメンに関するコンテンツを提示する、検索されたキーワードからメタデータを推定し、メタデータによる絞り込みを自動的に行うウェブ検索エンジンの実現を目指す。

1) Amanda Spink, Jack L. Xu. Selected results from a large study of Web searching: the Excite study. Information Research, Vol.6, No.1, 2000.
2) Karen Spärck Jones, Computer Laboratory. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, Vol.28, No.1, pp.11-21, 1972.
3) Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.
4) 総務省情報通信政策研究所. インターネット検索エンジンの現状と市場規模等に関する調査研究. 2009.

申請者氏名　　吉田光男

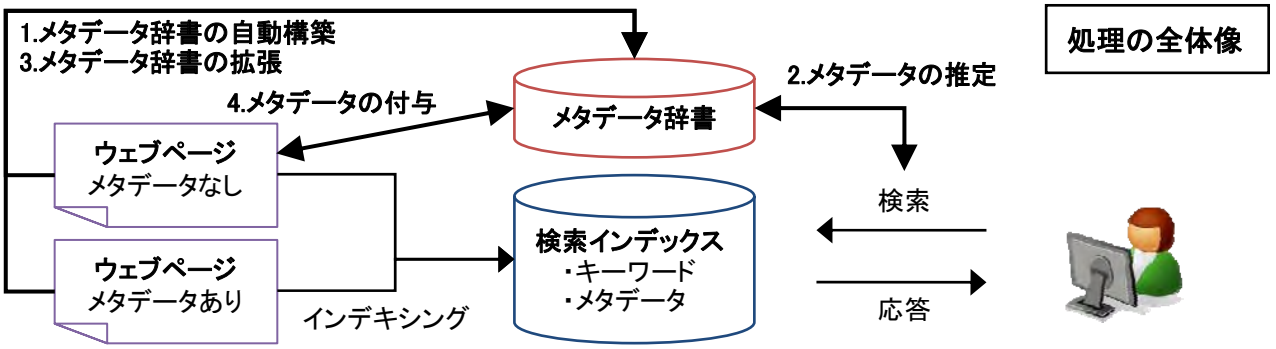
DC

(2) 研究目的・内容（図表を含めてもよいので、わかりやすく記述してください。）

① 研究目的、研究方法、研究内容について記述してください。
② どのような計画で、何を、どこまで明らかにしようとするのか、具体的に記入してください。
③ 共同研究の場合には、申請者が担当する部分を明らかにしてください。
④ 研究計画の期間中に異なった研究機関（外国の研究機関等を含む。）において研究に従事することを予定している場合はその旨を記載してください。

本研究の目的は、検索されたキーワードからメタデータを推定し、メタデータによる絞り込みを自動的に行うウェブ検索エンジンを実現する事である。時間や場所を限定した検索への高い需要が見込まれる、ブログに絞り込んで研究活動を進め、3 年後の実用化を目指す。なお、申請者は、大規模なウェブ検索エンジンを構築可能な研究チームとの共同研究を積極的に模索する。

本研究の対象とするメタデータは、時間「1270047600（UnixTime）」場所「緯度 35.6591 経度 139.7006（WGS84）」のような数値情報である。メタデータの曖昧性を排除し、実用化に近づける。以下、位置情報（場所、WGS84）を中心に、提案する 4 処理と評価方法について述べる。



1. メタデータ辞書の自動構築

キーワードからメタデータを推定するためには、「ハチ公 ⇔ 緯度 35.6591 経度 139.7006（WGS84）」のようにキーワードとメタデータが対となるメタデータ辞書が必要である。このような辞書は、従来、人手により作成されてきた。しかし、メタデータを持つ新語の存在があり、人手による辞書の作成には限界がある。そのため、メタデータ辞書の自動構築を検討する必要がある。

申請者は、Twitter をはじめとするマイクロブログのコンテンツを利用すれば、メタデータ辞書の自動構築が可能であると考えている。マイクロブログには、①投稿内容が短いため 1 トピックの傾向がある、②投稿の際に位置情報を付与できる、③事象発生から投稿までの間隔が短い、という特徴がある。この特徴を踏まえ、投稿内容に出現するキーワードと位置情報の対応付けを行う。

2. メタデータの推定

検索キーワードにメタデータが付与されているかどうかは、(1)で構築した辞書を参照することでわかる。しかし、1 キーワードに複数のメタデータが付与されている場合を想定する必要がある。今のところ、全てのメタデータで OR 検索を行う事と、最頻値となるメタデータのみで検索する 2 モデルを検討している。メタデータの推定方法は検索性能（利用者の満足度）に影響するため、(5)の評価のフィードバックを得ながら検討したいと考えている。

3. メタデータ辞書の拡張（同義語辞書の構築）

急増しているマイクロブログのコンテンツを利用してメタデータ辞書を自動構築すれば、様々なキーワードに対応できると考えているが、対応キーワードをより増やすために、ウェブページのコンテンツを利用して拡張を行う。住所表現を基に拡張、同義語辞書を構築して拡張の 2 通りを想定している。

住所表現による拡張では、ウェブページ中において、数値情報に変換可能な住所表現と、共に出現するキーワードとの対応付けを行う。正確に対応付けを行うためにはコンテンツ位置の識別が必要であるが、これまで研究してきたコンテンツの自動抽出手法を用いることで解決する。さらに、同義語辞書による拡張では、今日まで、様々な同義語辞書構築手法が提案されており、先行研究に倣い同義語辞書を構築する。同義語と(1)で構築した辞書のキーワードとを対応付け、辞書のサイズを大きくする。

4. ウェブページに対するメタデータの付与

メタデータによる絞り込みを自動で行うためには、ウェブページにメタデータが付与されていることが前提である。言語やファイルタイプなどは機械的に付与できるが、明示的に付与されていない地理的位置情報を推定するのは難しい。住所表現が含まれる場合は、(3)と同様の処理で解決できる。含まれていない場合は、ウェブコンテンツに出現するキーワードそれぞれのメタデータを推定し、その最頻値を求めることで解決する。

5. 実運用による評価（ブログ検索エンジンの構築）

検索システムの評価方法は、テストコレクションを作成して比較する方法が代表的である。しかし、利用者のニーズが多様化している今日、少数によって作成されたテストコレクションを用いる事は適切でないと考えられる。構築したブログ検索エンジンを利用者に触らせ、提示ページのクリック率、滞在時間を調べる事により、本研究の有用性を評価する。

申請者氏名　　吉田光男

DC

(3) 研究の特色・独創的な点

次の項目について記載してください。
① これまでの先行研究等があれば、それらと比較して、本研究の特色、着眼点、独創的な点
② 国内外の関連する研究の中での当該研究の位置づけ、意義
③ 本研究が完成したとき予想されるインパクト及び将来の見通し

本研究と類似した先行研究は、ウェブ検索エンジンの検索クエリ拡張 [1]が挙げられるが、あるキーワードが含まれるウェブページを探す事に重点が置かれたままである。本研究では、メタデータなどキーワード以外の情報をキーワードで検索することに重点を置いた。また、メタデータ辞書の構築のアイデアに似た先行研究として土屋ら[2] があるが、発見的手法に基づくものであり、ウェブ上で日々増加する新語に対応するのは困難であると考えられる。

ウェブページの数が膨大になる一方、ウェブ検索エンジンはランキングアルゴリズムの改良に偏り、絞り込み検索機能は貧弱なままである。本研究の意義は、できるだけ少ないコストでメタデータ辞書を構築する点、メタデータの付与されていないウェブページにメタデータを付与する点、ウェブ検索エンジンにおけるメタデータの絞り込みをキーワード検索のみで実現する点である。

本研究が完成すれば、キーワード検索のあり方に一石を投じるとともに、利用者の利便性を飛躍的に高めることができる。さらに、GPS 搭載携帯電話の普及により、場に応じた情報を検索するシーンが増えてきたが、ウェブページに地理的位置情報を付与することが可能になるため、既存コンテンツを有効に活用できる。

有用性評価のため、実際にブログ検索エンジンを構築し運用する。利用者からのフィードバックを直接受ける事ができるため、今後、インターネット利用者のニーズを的確に捉えた研究が期待できる。

1) 大石哲也, 倉元俊介, 峯恒憲, 長谷川隆三, 藤田博, 越村三幸. 関連単語抽出アルゴリズムを用いた Web 検索クエリの生成. 電子情報通信学会論文誌 D, vol.J92-D, no.3, pp.281-292, 2009.
2) 土屋誠司, 奥村紀之, 渡部広一, 河岡司. 連想メカニズムを用いた時間判断手法の提案. 自然言語処理, vol.12, no.5, pp.111-129, 2005.

(4) 年次計画

（１年目）

1. メタデータ辞書の自動構築

2. メタデータの推定

まず、マイクロブログのコンテンツを収集するクローラを開発し、メタデータ辞書の自動構築に必要なデータを継続的に収集する仕組みを構築する。次に、メタデータ辞書の構築を行う。その際、自動構築された辞書が人手によって作成された辞書（国土交通省の「国土数値情報」など）とどの程度乖離があるかを調査する。最後に、収集したマイクロブログを検索する仕組みを構築し、検索されたキーワードからメタデータを推定し、メタデータによる絞り込みを自動的に行うマイクロブログ検索エンジンを構築する。このシステムを用いて、時空間メタデータ検索をキーワード検索に統合することの有用性を利用者に評価させる。

（２年目）

3. メタデータ辞書の拡張（同義語辞書の構築）

4. ウェブページに対するメタデータの付与

まず、ブログのコンテンツを収集するクローラを開発し、メタデータ辞書の拡張や実運用に必要なデータを継続的に収集する仕組みを構築する。次に、メタデータ辞書の拡張を行う。そして、収集したブログページに対してメタデータの付与を試みる。最後に、構築したメタデータ辞書・同義語辞書を検索する辞書検索システムを構築する。このシステムを用いて、辞書の有用性を利用者に評価させる。

（３年目）

5. 実運用による評価（ブログ検索エンジンの構築）

前期は、本研究の成果を導入したブログ検索エンジンの構築を行う。このシステムを用いて、本研究の有用性を利用者に評価させる。利用者によるフィードバックを受けながら、アルゴリズムの改良を試みる。

後期は、本研究の一般化を試みる。本計画では、対象とするメタデータを時空間情報に絞っているが、他のメタデータにも適用可能であるかどうかを検討する。また、システム評価のあり方を検討し、できるだけ少ないコストで利用者の実感に近い評価方法を模索する。

申請者氏名　　吉田光男

4. 研究業績（下記の項目について申請者が中心的な役割を果たしたもののみ項目に区分して記載してください。その際、通し番号を付すこととし、該当がない項目は「なし」と記載してください。申請者にアンダーラインを付してください。）

(1) 学術雑誌等（紀要・論文集等も含む）に発表した論文、著書（査読の有無を区分して記載してください。査読のある場合、印刷済及び採録決定済のものに限ります。査読中・投稿中のものは除く）

① 著者（申請者を含む全員の氏名を、論文と同一の順番で記載してください。）、題名、掲載誌名、発行所、巻号、pp 開始頁－最終頁、発行年をこの順で記入し、著者の所属・職については脚注に記載してください。

② 採録決定済のものについては、それを証明できるものを P.8 の後に添付してください。

(2) 学術雑誌等又は商業誌における解説、総説

(3) 国際会議における発表（口頭・ポスターの別、査読の有無を区分して記載してください。）

著者（申請者を含む全員の氏名を、論文等と同一の順番で記載してください。）、題名、発表した学会名、論文等の番号、場所、月・年を記載してください。発表者に○印を付してください。（発表予定のものは除く。ただし、発表申し込みが受理されたものは記載しても構いません。その場合は、それを証明できるものを P. 8 の後に添付してください。）

(4) 国内学会・シンポジウム等における発表

(3)と同様に記載してください。

(5) 特許等（申請中、公開中、取得を明記してください。ただし、申請中のもので詳細を記述できない場合は概要のみの記述で構いません。）

(6) その他（受賞歴等）

(1) 学術雑誌等（紀要・論文集等も含む）に発表した論文、著書

【査読あり】

1-1) 吉田光男[1], 山本幹雄[2]. 教師情報を必要としないニュースページ群からのコンテンツ自動抽出. 日本データベース学会論文誌, 日本データベース学会, vol.8, no.1, pp.29-34, 2009.

注：著者の所属・職（論文発表時）

1. 筑波大学大学院システム情報工学研究科・大学院生　2. 筑波大学大学院システム情報工学研究科・教授

(2) 学術雑誌等又は商業誌における解説、総説　なし

(3) 国際会議における発表　なし

(4) 国内学会・シンポジウム等における発表

【口頭・査読あり】

4-1) ○吉田光男, 乾孝司, 山本幹雄. CSS セレクタで表現されたコンテンツ抽出ルールの自動獲得. 楽天研究開発シンポジウム 2009 論文集, pp.7-10, 品川シーサイド楽天タワー, 2009 年 11 月.

【口頭・査読なし】

4-2) ○吉田光男, 山本幹雄. 教師情報を必要としない Web ページ群のコンテンツ自動抽出ツールの提案. 第 1 回データ工学と情報マネジメントに関するフォーラム, A8-4, ヤマハリゾートつま恋, 2009 年 3 月.

4-3) ○吉田光男, 山本幹雄. 教師情報を必要としない Web ページ群の主要コンテンツ自動抽出. 第 23 回人工知能学会全国大会, 2B3-1, サンポートホール高松, 2009 年 6 月.

4-4) ○吉田光男, 乾孝司, 山本幹雄. ブログ記事集合を用いたポストとコメントとの自動分離抽出手法の提案. 情報処理学会研究報告, Vol.2009-DBS-149, No.20, pp.1-8, 慶応義塾大学, 2009 年 11 月.

4-5) ○吉田光男, 乾孝司, 山本幹雄. リンクを含むつぶやきに着目した Twitter の分析. 第 2 回データ工学と情報マネジメントに関するフォーラム, 5A-1, 淡路夢舞台国際会議場, 2010 年 2 月.

4-6) ○吉田光男, 乾孝司, 山本幹雄. リンクを含むつぶやきを中心とした Twitter の分析. 第 17 回 Web インテリジェンスとインタラクション研究会, pp.33-34, 大阪大学, 2010 年 3 月.

【ポスター・査読なし】

4-7) ○吉田光男, 山本幹雄. 教師情報を必要としない Web ページ群のコンテンツ自動抽出ツールの提案. 第 1 回データ工学と情報マネジメントに関するフォーラム, A8-4, ヤマハリゾートつま恋, 2009 年 3 月.

4-8) ○吉田光男, 乾孝司, 山本幹雄. CSS セレクタで表現されたコンテンツ抽出ルールの自動獲得. Web とデータベースに関するフォーラム 2009, 楽天推薦-1, 慶応義塾大学, 2009 年 11 月.

4-9) 吉田光男, 乾孝司, ○山本幹雄. ブログ記事集合を用いたポストとコメントとの自動分離抽出手法の提案. Web とデータベースに関するフォーラム 2009, DBS-11, 慶応義塾大学, 2009 年 11 月.

4-10) ○吉田光男, 乾孝司, 山本幹雄. リンクを含むつぶやきに着目した Twitter の分析. 第 2 回データ工学と情報マネジメントに関するフォーラム, 5A-1, 淡路夢舞台国際会議場, 2010 年 2 月.

4-11) ○吉田光男, 乾孝司, 山本幹雄. ブログページ集合からのポスト及びコメントの自動抽出. 言語処理学会第 16 回年次大会発表論文集, pp.418-421, 東京大学, 2010 年 3 月.

(5) 特許等　なし

(6) その他

【表彰】

6-1) 吉田光男, 山本幹雄. 第 1 回データ工学と情報マネジメントに関するフォーラム, 優秀インタラクティブ賞, 2009 年 3 月.

6-2) 吉田光男, 乾孝司, 山本幹雄. 楽天研究開発シンポジウム 2009, 優秀論文賞, 2009 年 11 月.

申請者氏名　吉田光男

DC

５．自己評価

日本学術振興会特別研究員制度は、我が国の学術研究の将来を担う創造性に富んだ研究者の養成・確保に資することを目的としています。この目的に鑑み、申請者本人による自己評価を次の項目毎に記入してください。

① 研究職を志望する動機、目指す研究者像、自己の長所等

② 自己評価する上で、特に重要と思われる事項（特に優れた学業成績，受賞歴，飛び級入学，留学経験，特色ある学外活動など）

【研究職を志望する動機、目指す研究者像、自己の長所等】

研究者を志望する理由は、不便に慣れてしまっているなど、問題を問題として認識されていない事柄を解決していくことにひかれるからである。その上、多数のウェブサービスを開発してきて、開発と研究は裏表の関係になってきているように感じた。より良いサービスを開発・提供するためには、本質を見極めるための研究を必要としている。私は、より良いサービスの一端を担う研究者になる。

私は、応用研究を中心とする研究者を目指す。実用化されずに埋もれていく研究を、非常に残念に思う。例えば、近年、EC サイトにレコメンデーションシステムが導入される事例が急増しているが、関連する研究は十数年前に活発に行われていた。現在導入されているシステムは、十数年前に提案された非常にナイーブなシステムである。早期に実用化されていれば、サイト利用者による生のフィードバックを受け、より研究が盛んになった可能性もある。逆に、最新の研究成果が導入されれば、より高度なレコメンデーションが可能となるはずである。我が国においては、研究から実用までのパスが非常に弱いと考える。理由は様々であろうが、少なくとも、そのパスになろうとした研究者が少なかったことは事実であろう。私は、そのパスになるような研究者になる。実用化に際しては、分野を横断して取り組む必要もある。自身の専門分野に固執せず、他分野の研究者及び技術者と積極的に交流し、互いにフィードバックすること、促すことにより、科学技術の向上に寄与する。

自身のこれまでの経験を踏まえ、私の長所は、大きく分けて以下の 3 つとなる。

1. 自ら問題を設定・設計し、解決する能力を有している
   これまでに様々なウェブサービスを開発し運営してきた（詳細は後述）。その理由は、既存のサービスに不満があるからであり、その不満を解決するために行った。不満を解決するためには問題点を明らかにし、解決するだけの確かな創造力、技術力、実践力が必要となる。これまでの研究においても、指導教員と異なる研究テーマに自ら取り組み、実績を残すことができている。
2. 継続する能力を有している
   ウェブサービスの開発に見落とされがちな点として、継続してサービスを運営するということが挙げられる。開発したサービスの大半を継続して運営しており、最長で 8 年以上継続している。これまでの研究においても継続的に対外発表を行い、成果の公開、フィードバックの獲得に努めている。
3. 研究を事業化する基盤を有している
   4 年前に有限会社てっくてっくを設立し起業した。てっくてっくは、筑波大学発ベンチャー企業として認定されている。研究成果を社会還元する方法の一つとして実用化が期待されており、その基盤を有している。ただし、博士後期課程入学後は研究に専念し、将来的には、研究と実用までのパスを強化する立場になりたいと考えている。

【自己評価する上で、特に重要と思われる事項】

私は、研究などの成果をオープンにすることを強く意識しており、今までの発表論文及び実験プログラムをオープンにしている。さらに、この流れを社会的にも加速させるため、My Open Archive[1)] の一員として啓蒙活動などを行っている。My Open Archive には、研究者、エンジニア、事務職員、翻訳者など、様々なバックグラウンドを持つ者が参加しており、協調して活動を行っている。

研究成果をオープンにする一環として、最新の研究成果を研究者の方に伝えられるよう、学会などでの発表も多数行っている。これまでに従事していたウェブコンテンツ抽出に関する研究は、ウェブに関する基盤研究であり、多くの研究者に伝えたいと考えていた。短期間で集中して研究を行い、発表を行った結果、「第 1 回データ工学と情報マネジメントに関するフォーラム」において優秀インタラクティブ賞を、「楽天研究開発シンポジウム 2009」において優秀論文賞を受賞した。

自己の長所に挙げたとおり、これまでに様々なウェブサービスを開発し運営してきた。最も継続して運営しているサービスは、2002 年 8 月に開始した CEEK.JP[2)] という統合型メタ検索エンジンである。そして、現在、最も認知されているサービスは、2004 年 6 月に開始した CEEK.JP NEWS[3)] というロボット型ニュース検索エンジンである。CEEK.JP NEWS は、Google News 日本語版よりも早くサービスを開始するなど、我が国で最も歴史のあるロボット型ニュース検索エンジンである。個人で様々なサービスを提供していることに注目を浴び、日経 Click[4)] や ITmedia[5)] にインタビュー記事が掲載された。

1. http://www.myopenarchive.org/　2. http://www.ceek.jp/　3. http://news.ceek.jp/
4. 福光恵. ネット次世代の原石たち. 日経 Click 緊急復刊号（日経ベスト PC+デジタル, vol.12, no.4）, 2007.
5. http://bizmakoto.jp/bizid/articles/0701/30/news116.html

申請者氏名　吉田光男

DC