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　１日現在

機関番号：３２７１４
研究種目：基盤研究（C）
研究期間：2008～2010
課題番号：２０５００１０３
研究課題名（和文）ワークフローと差分情報収集を用いたウェブ上の分散情報の動的収集システムの研究開発

研究課題名（英文）Development of a system for dynamically collecting information distributed on the Web using workflow and extraction of difference information

研究代表者
速水　治夫（HAYAMI HARUO）
神奈川工科大学・情報学部・教授
研究者番号：９０３０８５３６

研究成果の概要（和文）：Web 上に分散する知識情報を有効活用できるようにするため，情報収集ワークフローに基づき利用者をナビゲートし，重複した情報を含む複数のページからオリジナルな追加情報を抽出するシステムの研究開発を行った．具体的には，国際会議への参加を支援するシステム，個人の行事や行動の段取り力を高めるワークフローシステム，検索履歴を利用して情報検索を支援するシステム，市民グループの活動の位置情報を自動収集するシステムなどを開発し，有効性を確かめた．

研究成果の概要（英文）：We studied a system to make an effective use of knowledge information distributed on the Web. Our system navigates users using workflow to collect information, and extract additional information from Web pages containing duplicate one. We developed some subsystems on the Web and confirmed their effectiveness during the three year study period. They are systems to support participation in international conferences, to enhance individual project work ability, to help information retrieval using search history, to automatically collect information about locations of volunteer activities of citizen groups, etc.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,100,000	330,000	1,430,000
2009 年度	1,100,000	330,000	1,430,000
2010 年度	1,200,000	360,000	1,560,000
総　計	3,400,000	1,020,000	4,420,000

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：Web，ワークフロー，差分情報，ナレッジ共有，国際会議，段取り，検索履歴，位置情報

１．研究開始当初の背景

Web 上では世界中のさまざまなサイトに種々雑多な情報が分散して蓄積され，日々新しい情報が発生している．しかし，多くの情報は先行情報と重複しており，その上に新たな情報が付加されていることが多い．このため，検索結果の上位の大半が同一の内容であることも多い．したがって，適宜更新され膨大な情報の中から特定の目的のために有益な知識情報を収集することが非常に困難となっている．

たとえば「海外旅行の準備」という目的（一連の行為）を考えてみると，飛行機の乗り継ぎ，現地通貨の両替方法，空港事情，現地の

最新情報などを順次調査する．これらの情報は個人のブログや観光情報サイト，旅行会社のサイトに分散しており，それらは動的であったり静的であったりさまざまである．さらに重複した内容の Web ページも数多くある．しかし，それらの中には重複した情報に加えて，貴重なオリジナル情報を持つものも多い．

２．研究の目的

本研究課題では，さまざまな Web サイトに分散している知識情報を，特定の目的のために利用時に収集するシステムの研究開発を目的とする．具体的には以下の 2 つのアプローチにより，Web 上に分散する知識情報を有効活用できるようにする．

(1)ある目的のために情報収集した手順や利用した検索キーワードは同じ目的を持った他の利用者に役立つと考えられる．本研究では，特定の目的で情報を収集する手順を検索キーワードとともに蓄積する．これを「情報収集ワークフロー」とし，同じ目的を持つ利用者が情報収集する際，それを利用し，利用者の情報収集の道筋をナビゲートする．さらに，利用者の検索履歴を記録し，優先すべきことや進捗，ステップ毎の履歴などを提示する．また，前のステップの履歴を利用者のバックグラウンドとして利用し，以降のステップのナビゲーションに活用する．

(2)検索結果において重複した情報が多くの Web ページに含まれる場合,それぞれの Web ページに含まれるオリジナルな追加情報にこそ価値があると考えられる．本研究では，検索結果からほぼ同一の内容を持つ複数のページを発見し，それらの差分を情報の記録された日時や参照関係を利用してダイナミックに抽出する．

本研究により，これまで以上に Web を知識源として有効活用できるようになり，また，膨大な情報から価値ある情報を見つけ出す可能性を高めることが期待できる．

３．研究の方法

本研究課題における研究方法の重要なポイントは，(a)情報収集ワークフローに基づき利用者をナビゲートする，および(b)重複した情報を含む複数のページからオリジナルな追加情報を抽出するために，(ⅰ)情報収集の手順と検索キーワードからなる情報収集ワークフローの設計・管理，(ⅱ)利用者の状況やバックグラウンドによる的確な情報収集のナビゲーション，(ⅲ)重複部分を特定し複数のページから差分情報のみの抽出，(ⅳ)情報収集ワークフローや差分情報の効率的な提示，これら 4 つの技術を確立し，実用化することにある．本研究課題では汎用的な仕組みの実現を目指すが，研究の初期段階では具体的なテーマとして国際会議への参加支援に絞る．

(1)平成 20 年度の研究方法

上記(ⅰ)から(ⅲ)の要素技術の基本方式を開発する．Web とその利用環境の現状や今後の技術革新を調査・検討した上で，各要素技術について基本方式を考案する．そして，それぞれのプロトタイプシステムの処理フローやデータ構造を設計・実装し，動作や性能を検証する．

(ⅰ)では，特定の目的のために必要と思われる一連の情報を利用者が簡単に検索できるように，効果的で基本的なキーワードを検索の手順に従って作成し管理する情報収集ワークフローの手法を開発する．(ⅱ)では，利用者のこれまでの検索履歴を活用し，利用者の状況やバックグラウンドと情報収集ワークフローを照合しながら，利用者がスムーズに情報収集できるようにナビゲートする手法を開発する．(ⅲ)では，Web ページ間の重複部分を特定し差分情報のみを上手に抽出する方法を開発する．そのため Web ページを構成する重要キーワードと出現箇所や関連性などの情報を有機的に組み合わせて活用することを検討する．(ⅳ)については，文献調査や学会・研究会などへの参加を通じ効果的なユーザインタフェースの検討を行う．

(2)平成 21 年度と平成 22 年度の研究方法

平成 21 年度は，平成 20 年度に整理した効果的な提示方法に基づき，システムのユーザインタフェースを設計し開発する．そのインタフェースに平成 20 年度に実現した情報収集ワークフローに基づくナビゲーション技術と差分情報の抽出技術を統合し，全体としてひとつのシステムとして開発する

平成 22 年度は，平成 21 年度に開発したシステムを実運用し有効性を検証する．そのため，システムを Web 上で公開し，アクセスログの収集やアンケート調査を行い，結果を分析する．

４．研究成果

本研究課題の期間中に，国際会議に参加することを支援するシステム，個人の日常活動の段取り力向上を目指したワークフローシステム，検索履歴を利用した Web ページを推薦するシステム，市民グループの活動の位置情報を自動収集するシステム，ネットショッピングに関する情報を一元的に管理するシステム，Wikipedia を情報源として質問応答するシステムなど，本研究課題の目的達成に向けて多種多様なシステムを開発した．ここでは代表的なシステムについて特徴や機能，評価結果を述べる．

(1)国際会議参加支援のためのナレッジ収集・共有システム

国際会議に参加する人，特に発表経験が少

ない人を支援するため，参加までの一連の作業で必要なナレッジを共有するシステムを開発した（図 1）．

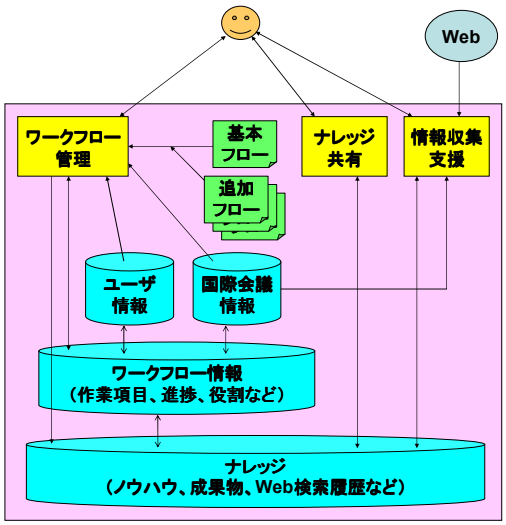

図 1　国際会議参加支援のための
ナレッジ収集・共有システムの構成

本システムでは，国際会議に参加するまでの一連の作業をワークフローと考え，ナレッジとワークフローを関連付けて管理する．さらに，投稿締切りなどのデッドラインや共著者は国際会議ごとに異なることを考え，本システムでは，ワークフローの生成時や進行にあわせて，作業のデッドラインや共著者の役割などを追加・修正する．ワークフロー管理機能により，発表者や共著者の作業を促進したり，進行状況を把握したりしやすくする．またナレッジが作業単位で整理される．

さらに本システムでは，Web も重要な情報源になっていることを踏まえ，検索エンジンに入力したキーワードも再利用の可能性が高いナレッジとして共有する．具体的には，Web で情報を調べる場合，一般に検索エンジンを利用する．しかし国際会議の初心者にとって何を調べておくべきかわからない．一方，過去に検索した結果は時間の経過とともに情報が古くなってしまう可能性があるが，そのときに利用した検索キーワードは将来でも役立つ可能性がある．そこで本システムでは，利用者が検索するときに利用した検索キーワードを検索履歴として蓄積し，国際会議に参加するためのワークフローと関連付けて管理する．国際会議のワークフローと検索履歴を時系列順に閲覧可能にすることで，利用者の情報収集をナビゲートする．

我々の 1 人が本研究課題の期間中に参加した 2 つの国際会議のために本システムを利用し，同時に，これらの会議に参加するためにいつ何を実施したかも記録した（以下，作業記録）．国際会議に参加するとき，過去に参加したときの投稿論文や各種書類を参考にすることがよくある．本システムでは，各作業で蓄積された情報はその作業に関連付けられるため，上記のような状況に効果がある．Web での情報収集のために利用された検索キーワードに注目すると，論文投稿までの作業では，どちらの会議においても，国際会議の Web サイトを訪れ論文のフォーマットを確認したり，所定のフォーマットにしたがって論文を作成するためのテクニックを調べたりするための検索キーワードが多く利用されていた．また，利用された検索キーワードと作業記録を照合すると，それらの内容からどの作業を実施しようとしていたのかがある程度理解できた．システムに蓄積された検索キーワードを国際会議のスケジュールにあわせて，適切なタイミングで提示することができれば，作業の実施時期や何を調べておくと良いのかを示すことができ，国際会議の経験がほとんどない人に有効であると思われる．国際会議に合わせて検索キーワードを展開する技術の開発などが課題である．

(2) 個人の行事や行動の段取り力を高めるワークフローシステム

個人の仕事，行事や行動の作業を考えた場合，効率よく進める人とそうでない人がいる．これらの作業を効率よく進めるチカラは”段取り力”と呼ばれ，多くの人はその向上に努めている．段取り力の向上には，自己の経験の蓄積に加え，優れた段取り力を持った人の経験に学ぶことが重要である．段取りの作成・実行に，企業の業務支援で成功しているワークフローシステムが考えられる．本研究課題では，個人の段取り力向上が可能となるワークフローシステムを開発した（図 2）．本システムはスケジューラと段取りデータが連動する．

図 2　段取り力を高める
ワークフローシステムの構成

本システムでは段取りは時系列順に発生する，または並列作業可能な work-item から構成され，各 work-item には作業内容，タグ（キーワード），コメントなどを付与することができる．ユーザは段取りが対象とする予

定の日時を切り離した段取りクラスを蓄積する（段取りリスト）．また本システムは，利用者がスケジューラに予定している物事を入力すると，それに適した段取りクラスを推薦する．推薦アルゴリズムは，(1)過去の予定から推薦，(2)利用者ごとに管理される段取りリストから推薦，(3)最近使用した段取りから推薦，という 3 ステップを経る．

本システムで段取りクラスを作成する際は，段取りクラスを新規作成するか，他の利用者の段取りクラスをコピー・アレンジして作成する．利用者がスケジューラに予定を入力すると，上述の推薦アルゴリズムをとおして段取りクラスが適用され，段取りインスタンスが生成される．また，予定を確認するたびに段取りインスタンスの残り作業も同時に確認することができる．段取りの各 work-item にはさまざまなメモが残されている．これらの仕組みによりいつ何をやるべきかが明確となり，それにより必要な情報の収集も行いやすくなる．

情報系学部・大学院の学生 6 名に開発システムの実験に協力していただき評価を行った．その結果，他の利用者の段取りクラスをコピーする機能や，予定を入力すると段取りクラスが推薦される機能は極めて高い評価を得た．また，推薦された段取りクラスの妥当性も良い結果であった．一方，入力された予定から段取りクラスを推測しにくい場合，利用者自身の段取りリストのみを利用して適切な段取りクラスを推薦することが困難であることがわかった．また，ユーザインタフェースの向上が必要であることもわかった．

(3)検索履歴の共有による情報検索支援システム

「同じ目的で検索した先人は必ず存在する」という観点から，あるユーザの Web 検索履歴を，検索キーワード，閲覧ページ，検索キーワードの遷移により定式化し，同一目的の別のユーザの検索を支援するシステムを開発した．本システムは検索履歴を蓄積するためのモジュールと，それを利用して検索者の検索目的を推定し Web ページを推薦するためのモジュールから構成される．

検索履歴蓄積モジュールでは「Web ページ内の文字列をコピーし検索エンジンの入力フォームにペーストして新たに検索する」という，検索プロセス内で頻繁に行われると思われる操作に着目した．本システムでは，ある検索キーワードを利用して検索した結果から閲覧した全ての Web ページ（Web ページ群）をその検索キーワードとともに「検索ブロック」として定義し，さらに，Web ページ群のひとつのページに含まれる文字列を新たな検索キーワードとして検索した場合に，その文字列を「リンク文字列」と定義する．本システムは，検索ブロック間をリンク文字列で関連付けて蓄積することにより「検索キーワードを入力して検索し，検索結果から Web ページをたどり，検索キーワードを改めて再検索し，最終的に目的の情報を得る」という一連の検索プロセスを管理する．また，最終的に目的の情報が得られた場合，検索履歴の蓄積者はそのことを示すためにマークすることができる．

Web ページ推薦モジュールは，利用者（検索者）が入力した検索キーワードと検索結果に含まれる Web ページ内で次の検索キーワードとして選択されたリンク文字列の対（検索キーワード対）と，検索履歴蓄積モジュールにより蓄積された検索履歴中の検索キーワードとリンク文字列の対とを比較し，検索者の検索キーワード対に合致する検索履歴が存在した場合，検索者と検索履歴の蓄積者の検索目的が同一であると判断し，その検索プロセスにおいて最終目的の情報としてマークされた Web ページを推薦する．

本研究課題では，両モジュールに必要なサーバサイドのプログラムと Web ブラウザ（本研究課題では Firefox を対象とした）の拡張機能を開発し（図 3），情報系学部・大学院の学生 10 名の協力を得て評価した．その結果，検索履歴の蓄積方法は操作に慣れれば問題なく実施できるものの，有益な Web ページを推薦するためには数多くの検索プロセスが必要であり，それをいかにして収集・蓄積するのかが課題であると認識した．検索履歴蓄積モジュールの操作については肯定的な回答であったため，継続的にシステムを利用し，検索履歴の数を増やした上での評価が必要である．

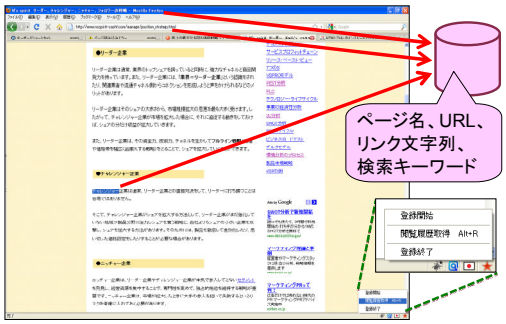

図 3　検索履歴モジュールの概要図

(4)市民グループの活動の位置情報の自動収集システム

市民グループ（NPO 法人や任意のボランティア団体）はそれぞれのグループが思い思いに Web サイトを立ち上げて情報発信しているため，ボランティア活動に参加したいという意欲のある人などが，どこでどのような活動が行われているかを理解することが困難で

ある．そこで，市民グループ（特に NPO 法人）の Web サイトや NPO 法人のオンラインデータベースから位置情報（住所）を自動的に収集しデータベース化するシステムを開発した．

本システムでは，まず初めに，(1)NPO 法人のオンラインデータベースから各 NPO 法人の名称や所在地，活動分野・目的などの基本情報を取得し，所在地の緯度経度を付与して基本情報データベースに追加する．次に，(2)検索エンジンを利用して，各 NPO 法人の Web サイトのアドレス（URL）を収集・選択する．次に，(3)NPO 法人の Web サイトから住所と地図のメタデータ（地図が利用されている Web ページの URL や，地図画像のファイル名，地図情報サービスへのリンク）を抽出し，住所には緯度経度を付与して一次データベースに追加する．そして，(4)抽出された住所について，Web ページの構造や地図の有無などを利用して住所のフィルタリングを行い二次データベースに蓄積する．そして，(5)フィルタリング後の住所を地図情報サービスを利用して地図上に表示する（図 4）．

本システムの住所のフィルタリングでは，Web サイトから抽出された住所のいずれも，基本情報データベースに蓄積されている所在地と一致しない場合，その Web サイトは誤って選択されたと見なし，全ての住所を二次データベースに追加しない．次に，各 NPO 法人の Web サイトについて，XML 文書として読み込んだ HTML のツリー構造上の同じ位置に何度も記載されている住所は，Web サイトに統一のメニューやヘッダー，フッター中に存在すると考えられる．そのため，同一ページ内で，メニューなどの位置以外に住所が記載されている場合は，メニューなどに記載されている住所を二次データベースに追加しない．また，Web サイト上の複数の異なるページにおいて，ツリー構造上の同じ位置に異なる住所が記載されている場合，検索された Web サイトがオンラインデータベースのひとつである可能性がある．そのため，その位置に対象となっている NPO 法人の所在地が記載されている場合，その Web サイトからは住所を二次データベースに追加しない．

本研究課題では神奈川県内に所在地を持つ NPO 法人を対象としてシステムを試作し，Web サイトの選択や住所の抽出精度を評価した．その結果，Web サイト選択の適合率は 75%近かったが，再現率が約 6 割であり，再現率の向上が課題となった．一方，住所の抽出精度について，本システムのフィルタリング手法を利用すると，間違った住所が抽出されることが少なかったが，もう少し多くの住所を抽出できるように工夫する必要がある．市民グループの活動を支援する立場の人からの意見では，システムの有効性を示す意見が多数得られた．なお，本システムは現在，神奈川県内の NPO 法人の協力を得て，Web 上で公開されている．

図４　市民グループの活動の位置情報の地図表示例

これら 4 つのシステムを「３．研究の方法」で述べた 4 つの要素技術に対照させると，(1)と(2)のシステムでは「(ⅰ)情報収集の手順と検索キーワードからなる情報収集ワークフローの設計・管理」と「(ⅱ)利用者の状況やバックグラウンドによる的確な情報収集のナビゲーション」を実現し，(3)のシステムでは(ⅱ)を実現した．(4)のシステムでは「(ⅲ)重複部分を特定し複数のページから差分情報のみの抽出」を実現した．また本研究課題で開発した全てのシステムにおいて「(ⅳ)情報収集ワークフローや差分情報の効率的な提示」を実現している．これらを満たす研究は我々報告者の知る範囲では行われておらず，極めてオリジナリティが高い研究である．

現在，twitter や facebook などソーシャルメディアと呼ばれるサービスが爆発的に人気を得ており，Web 上にはこれまで以上に種々雑多な情報が数多く存在する．またそれらの新しいメディアは相互に連携しており，同一の情報がさまざまなサービスに分散して存在していたり，新しい情報が次々と追加されて蓄積されている．そのような中から有益な情報を見つけ出すことは大きな課題となっており，本研究課題の成果は今後，ソーシャルグラフと呼ばれる人間関係を利用したりして，新しいメディアにおける情報検索の支援に発展させることが期待できる．

５．主な発表論文等

〔雑誌論文〕（計２３件）

① Hitomi Sato，Akira Hattori，Haruo Hayami：A System to Share Arrangements for Daily Tasks and Life Events on the Web，Proceedings of 14th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems，査読有，2010，pp.290-297

② 佐藤仁美，服部哲，速水治夫：日常作業やイベントの作業効率向上に向けた段取り共有システムの提案，マルチメディア，分散，協調とモバイル（以下ではDICOMO）（DICOMO2010）シンポジウム論文集，査読有，2010，pp.1381-1386
③ 服部哲，五百蔵重典，速水治夫：市民活動団体のWebサイト上の活動場所情報の自動収集システム，DICOMO2010シンポジウム論文集，査読有，2010，pp.677-682
④ 石田武久，服部哲，速水治夫：閲覧履歴の共有による情報検索支援システムの提案，DICOMO2010シンポジウム論文集，査読有，2010，pp.414-419
⑤ Akira Hattori，Shigenori Ioroi，Haruo Hayami：Web-Based System for Supporting Participation in International Conferences，Proceedings of 13th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems，査読有，2009，pp.631-638
⑥ 服部哲，五百蔵重典，速水治夫:ナレッジ共有に基づく国際会議参加支援システム，DICOMO2009シンポジウム論文集，査読有，2009，pp.1211-1216
⑦ 佐藤仁美，服部哲，速水治夫：情報の関連性を重視したWebスケジューラ，DICOMO2009シンポジウム論文集，査読有，2009，pp.1206-1210
⑧ 速水治夫，渡邊裕一，服部哲，五百蔵重典：Webサイトアクセスログの時間的・地理的分析システムの提案，DICOMO2009シンポジウム論文集，査読有，2009，pp.749-755
⑨ 服部哲，小林直矢，速水治夫：音楽情報サイトからキーワードを自動取得するWeb楽曲データベースの提案，DICOMO2008シンポジウム論文集，査読有，2008，pp.1037-1040
⑩ 服部哲，速水治夫：Webを活用した国際会議発表支援のためのナレッジ収集・共有システム，DICOMO2008シンポジウム論文集，査読有，2008，pp.1041-1046
⑪ 三枝優一，服部哲，速水治夫，奥村学：Wikipediaの語彙資源を利用したインタラクティブ型Web質問応答システム，DICOMO2008シンポジウム論文集，査読有，2008，pp.1793-1802

〔学会発表〕（計１６件）
① 渡邉岳志：楽曲のキーワードの類似度を用いたプレイリスト作成支援システム，情報処理学会研究報告，2011.03.18，熊本
② 芳賀優人：行動ターゲティングを利用した行き先提示システム，情報処理学会研究報告，2011.03.17，熊本
③ 渡邊裕一：自然言語からの属性情報を考慮したオブジェクト抽出手法に関する検討，第9回情報科学技術フォーラム講演論文集，2010.9.7，福岡
④ 服部哲:市民活動情報の共有のためのWebベースマップシステムの検討，2010年日本社会情報学会（JASI&JSIS）合同研究大会論文集，2010.9.5，長崎
⑤ 服部哲：Webを情報源とした市民グループの活動場所の自動収集システムの提案と試作，電子情報通信学会　第15回サイバーワールド研究会報告集，2010.3.10，東京
⑥ 勝井美紗緒：商品情報の一元管理によるネットショッピング支援システム，情報処理学会研究報告，2010.1.21，神奈川
⑦ 服部哲：地域的広がりと時間的変遷のある地域情報を管理可能なシステムの提案，2009年日本社会情報学会（JSIS&JASI）合同研究大会論文集，2009.9.12，新潟
⑧ 佐藤仁美：複数のビューで提示可能な協調作業管理システムの提案，情報処理学会研究報告，2009.3.18，神奈川

〔図書〕（計３件）
① 速水治夫，納富一宏：データベースの実装とシステム運用管理，コロナ社，2010，146ページ
② 速水治夫，古井陽之助，服部哲：Webデータベースの構築技術，コロナ社，2009，187ページ
③ 速水治夫：リレーショナルデータベースの構築技術，コロナ社，2008，153ページ

６．研究組織
(1)研究代表者
速水　治夫（HAYAMI Haruo）
神奈川工科大学・情報学部・教授
研究者番号：90308536
(2)研究分担者
五百蔵　重典（IOROI Shigenori）
神奈川工科大学・情報学部・准教授
研究者番号：20318992
服部　哲（HATTORI Akira）
神奈川工科大学・情報学部・准教授
研究者番号：60387082