科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　２　日現在

機関番号：　１２１０２
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００２４５

研究課題名（和文）形式概念分析によるユーザーの直観に適したウェブ検索結果の視覚化システムの開発

研究課題名（英文）Visualization systems based on formal concept analysis and their application to web retrieval

研究代表者
延原　肇（Nobuhara, Hajime）

筑波大学・システム情報系・准教授

研究者番号：８０３５９６８７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円、（間接経費）　　960,000円

研究成果の概要（和文）：本プロジェクトでは、ウェブ検索結果をはじめとする、ゲノム、画像群、医療データ等の大規模な情報を、人間の直観にわかりやすい形式の一つとしての概念構造として視覚化する方法として、形式概念分析に基づく手法を提案した。特に、ウェブ検索結果およびゲノムの視覚化においては、関連識者による評価を通して、提案手法の有効性を確認している。本プロジェクトの遂行により、雑誌論文7編、国際会議および国内学会での発表多数、特許出願1件、という業績を挙げている。

研究成果の概要（英文）：Based on using formal concept analysis, visualization systems are proposed for the web retrieval results, genome array information, image retrieval results, medical information, as the big data. In this project, especially, with respect to web retrieval results, genome array information, it has been confirmed the effectiveness of the proposed method by various expert persons. As the publications, we obtained 1 patent, 7 journal papers (including international journal) and many conference/domestic papers.

研究分野：　計算知能

科研費の分科・細目：　情報学

キーワード：　形式概念分析　情報視覚化　情報可視化　ウェブ・インテリジェンス　ビッグデータ

１．研究開始当初の背景

Web 上の 1 兆を超える URL，さらにそれを利用するユーザーも増加傾向にあり，日本だけでも 9400 万人，全世界では，2013 年に 22 億人を突破すると予測されている.これらネット利用者の約 80%は，膨大な量の情報がインターネット上に存在するにも関わらず，1 回あたりの検索において閲覧するのは「わずか 3 ページ」，それらのページに目的の情報を見つけられない場合は，新規クエリによる再検索のプロセスに入ってしまう．このような従来の検索プロセスを効率化することができれば，ネット利用者数が膨大であるがゆえ，計り知れないほどの経済的効果，社会的な恩恵を生み出すことができる．

現行の検索エンジンの問題点は，ユーザの入力クエリに対して，タイトル，URL，及び概要文をランキング形式で出力する点にある．このランキング形式では，上位から順次クリックして辿ってゆくため 3 ページ以内に目的情報が得られない場合，再検索の作業に戻ってしまう．また，ランキング形式では各ページの相互関係が不明確のため，上位の閲覧済みのページ情報等が，下位の未閲覧のページの検索絞り込みに対して役に立たない．ランキングされている各ページの概要文は，クエリを含む文章を断片的に出力するだけで，そのページの本質を反映しているとは言えず，ユーザーの咀嚼しやすい直観的な形式に抽象化されているとはいえない．

２．研究の目的

本研究の目的は，従来の検索エンジンの問題点を解決するための視覚化システムを開発することである．以下，提案システムのポイントを記述する．

提案システムでは，検索結果をランキング形式ではなく，人間の直観に適した形式，すなわち各ページが含むキーワードの意味の広狭の観点から概念の階層構造(半順序関係)として提供する．この概念構成には申請者が長年研究している形式概念分析を用いる．

この概念の階層構造では，上位概念にあるページが下位概念にあるページとリンクする形で視覚化されている．普段の人間の検索行為においても，大局的な観点から詳細な局所へ視点を移動させており，提案システムが提供する情報構造は，人間の直観的な検索に適しており，上位概念の閲覧履歴を用いて，下位概念にあるページを絞りこむことで，検索効率が大幅に向上する．

提案システムでは，ウェブという非構造なテキストデータから，共起頻度およびウェブ独自の特性であるリンク構造を用いて，そのページを本質的に表す重要なキーワードを的確に抽出することで，適切な概念構造の構成を行う．

以上の問題解決のアプローチを用いることで，ユーザは従来の検索エンジンの出力結果を猛禽類のように探索することができ，検索効率の向上が実現できる．

３．研究の方法

申請者は，本研究の中核となる視覚化システムのプロトタイプ「Search@once」の開発を完了しており，国内外の学会において成果を報告，さらに海外の基調講演に招待される等，注目を集めている．本研究の研究期間内では，このプロトタイプシステムをベースに，3 つの観点から改良を行う．

提案システムおよびそのベースとなるプロトタイプシステムでは，形式概念分析を用いて，検索結果のウェブページの半順序関係を形成する．ここでポイントとなるのは，ウェブに含まれる膨大なキーワードから重要なキーワードを的確に抽出することであり，本研究では，ウェブ独自のリンク構造等を利用した本質的なキーワード抽出法を新規提案する．

現在のプロトタイプシステムでは，形式概念分析で得られた階層構造をそのまま視覚化しているが，よりユーザーの直観に適した検索結果を提示するためには，各階層構造に対して適切なラベルを割り当て，見やすい視覚化を行う必要がある．本研究では，ウェブという膨大な外部リソース，特にウィキペディアのようなカテゴリ情報を持つ集合知の情報源を適切に利用することで，各階層に適切なラベルを付与する手法を開発する．

現在のプロトタイプシステム「Search@once」では，YahooAPI 等をオンラインで利用しているため，レスポンスが遅くなってしまう欠点がある．検索結果をローカルにキャッシュし，検索時間を高速にするといった技術的な向上にも挑戦する．

４．研究成果

研究開始１年目で得られた成果は、形式概念分析に基づいた検索エンジンの視覚化技術のプロトタイプを Python に基づき開発を行ったことである。具体的なシステムのプロセスとしては、ユーザーの任意の検索クエリに対して，Yahoo API を用いて検索結果のテキストページ群を取得する．取得後のテキストページを，Beutiful Soup および形態素解析 API を利用することでキーワードに分割する．さらに，それらのキーワードの重要度，頻度を計量した上で Web ページおよび対応するキーワード群の関係表を作成する．ここで各キーワードの重要度の計量には，篠崎らの

提案している手法「リンク構造と語の共起を利用したWebページからのキーワード抽出手法とページ要約による検索支援」（第3回 楽天研究開発シンポジウム）を採用した．構成した情報表に対して形式概念分析を適用し，検索結果のwebページ群を概念構造の観点から抽象化し，ユーザーに対して咀嚼しやすいシステムを開発している．

また、従来の形式概念分析では，得られた概念構造が複雑になりすぎて具体的に出力された結果が何を示しているのかよくわからない，といった現象を引き起こしていたが，本研究で提案する束構造視覚化では，力学系モデルを採用し，また束構造をできるだけ保持した形で視覚化することで，この問題を解決している．この年度の大きな成果として，国内特許出願（情報検索支援装置，特願2012-2539）が1件，また国際会議1件，国内学会（査読つき）1件が得られている．

2年目で得られた成果としては、形式概念分析に基づいた直感的なインタフェースとして，2つの大きな応用事例を開発し，さらにその拡張として，学習支援への応用も行ったことである．それぞれを以下に示す．

まず応用事例の1つめとして，Web上の画像検索を対象としたシステムの開発を行った．具体的なシステムの流れとしては，ユーザーの任意の検索クエリに対して，Google画像検索を用いて検索結果の画像および周辺テキストページ群を取得する．取得後のテキストページを，形態素解析API等を利用することでキーワードに分割する．さらに，それらのキーワードの重要度，頻度を計量した上で画像および対応するキーワード群の関係表を作成し，情報表に対して形式概念分析を適用し，検索結果の画像群を概念構造の観点から抽象化し，ユーザーに対して咀嚼しやすいシステムを開発している． さらに，前年度に開発した力学系モデルに基づく束構造視覚化を用いて，画像検索結果の視覚化もより咀嚼しやすいものに対応させている．

応用事例の2つめとして，動画閲覧サイトYoutubeを対象とした，動画推薦システムを形式概念分析を利用して開発した．従来の動画推薦では，推薦理由が明示されない点を，形式概念分析の概念の連結情報を利用し，推薦理由を明示するシステムを構築した．

最終年度の成果として、これまでに得られた知見を最大限に活用し、3つの応用領域における試験的展開を行った。

まず応用事例の1つめとして、テキストデータ、特にインターネット上に公開されている電子テキストを対象に形式概念分析を適用し、その視覚化を行った。この応用の枠組みでは、国会議事録をはじめニュースサイトなどへの適用を行い、国会議事録の視覚化においては、東日本大震災前後において重点的に議論されていた項目等が、適切な情報粒度で抽出することができることを確認した。

応用事例の2つめとして、農業分野における育種支援のツールとして、形式概念分析を利用したゲノム配列の視覚化を行った。これに関しては、実際にシステムの実装を行い、この結果を育種の専門家に評価してもらい高い評価を得るとともに、その成果は国際雑誌論文に採録されている。

応用事例の3つめとして、医療診断における診断支援ツールとして、形式概念分析とニューラルネットワークを組み合わせたシステムを開発し、それを実際の医療データに適用した。この研究成果については、国際会議等において報告している。

以上、本研究では、これまでの3年間の研究を通して、形式概念分析を利用した (1) 検索結果の視覚化、(2) 動画検索結果の視覚化および推薦理由の明示、(3) 画像検索結果の視覚化、(4) 電子テキスト（特に議事録）の視覚化、(5) 育種支援のためのゲノム配列の視覚化、(6) 医療診断支援のための視覚化、という6つの多様な分野への応用を行い、それぞれの分野において高い評価を得るとともに、具体的な業績として多くの雑誌論文および国際会議において採録されており、本研究はこれらの観点から十分な成果が得られていると言える。

５．主な発表論文等

〔雑誌論文〕（計 8 件）

(1) H. Hashikami, T. Tanabata, F. Hirose, N. Hasanah, K. Sawase, and H. Nobuhara: An Algorithm for Recomputing Concepts in Microarray Data Analysis by Biological Lattice, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vo. 17, No. 5, pp. 761-771 (2013)

(2) K. Sawase and H. Nobuhara: The transformation method between tree and lattice for file management system, Evolving Systems, 10.1007/s12530-012-9071-4 (2013)

(3) 北村 祐太郎, 澤勢 一史, 延原 肇: 形式概念分析を用いた推薦理由を明示する動画推薦手法, 日本知能情報ファジィ学会誌, Vol. 25, No. 1, pp. 624-635 (2013)

(4) 豊田 哲也, 延原 肇: カテゴリ写像に基づく追加学習に対応可能な自己組織化とWebニュース群の動的クラスタリングへの応用, 電気学会論文誌 C, Vol. 132, No. 8, pp. 1347-1355 (2012)

(5) T. Tanabata, F. Hirose, H. Hashikami, and H. Nobuhara: ‘Interactive Data Mining Tool for Microarray Data Analysis Using Formal Concept Analysis,’ Journal of Advanced Computational Intelligence and

Intelligent Informatics, Vol. 16, No. 3, pp. 273-281 (2012)
(6) S. Kawauchi, and H. Nobuhara：’ Knowledge Expansion Support by Related Search Keyword Generation Based on Wikipedia Category and Pointwise Mutual Information,’ Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 16, No. 3, pp. 247-255 (2012)
(7) T. Sugimoto, and H. Nobuhara：‘A Recommendation System with the Use of Comprehensive Trend Indication Based on Weighted Complete Graph’ Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 16, No. 3, pp. 266-272 (2012)
(8) T. Toyota, and H. Nobuhara：’ Visualization of theWeb News based on Efficient Self-Organizing Map using Restricted Region Search and Dimensionality Reduction’ Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 16, No. 3, pp. 219-226 (2012)

〔学会発表〕(計 7 件)
(1) 澤勢一史，延原 肇，形式概念分析を用いた国会議事録の束構造可視化と時空間解析，電子情報通信学会 スマートインフォメディアシステム研究会，クリエート浜松，静岡県、 2013 年 3 月 7-8 日
(2) 澤勢一史，延原 肇，周辺テキストおよび感性情報に基づく形式概念分析を用いた画像群の概念構造化，第 28 回ファジィシステムシンポジウム，名古屋工業大学，2012 年 9 月 12 日-14 日
(3) 吉永直嗣，延原肇，ウェブのリンク構造と語の共起を利用したキーワード抽出に基づく情報検索結果の概念構造化，WebDB Forum 2011，Web とデータベースに関するフォーラム，工学院大学，2011 年 11 月 4 日
(4) H. Hashikami, N. Hasanah, K. Sawase, T. Tanabata, F. Hirose, and H. Nobuhara,‘An Efficient Recomputing Concepts Algorithm for Microarray Data Analysis Using Biological Lattice,’ 2011 International Workshop on Smart Info-Media Systems in Asia (SISA2011), Nagasaki, Japan, Oct. 31 - Nov. 2, (2011)
(5) 澤勢一史，延原肇，概念の接続情報に基づく束構造の複雑さの定量化，第 27 回ファジィシステムシンポジウム，福井大学，2011 年 9 月 12 日-14 日
(6) N. Yoshinaga, A. Shinozaki, H. Nobuhara,‘Formal Concept Analysis Based Information Retrieval Support System with Keyword Extraction using Web Links and Word Co-occurence,’ The 2011 IFSA World Congress and the 2011 AFSS, Surabaya, Indonesia, Jun. 21, (2011)
(7) H. Nur, K. Sawase, H. Nobuhara,‘Mining Rules for Diagnosis of Dengue Hemorrhagic Fever using Neural Networks and Formal Concept Analysis,’ The 2011 IFSA World Congress and the 2011 AFSS, Surabaya, Indonesia, Jun. 21, (2011)

〔産業財産権〕
○出願状況（計 1 件）

名称：情報検索支援装置
発明者：延原肇、吉永直嗣、澤勢一史
権利者：同上
種類：特許
番号：特願 2012-2539 号
出願年月日 平成 24 年 1 月 10 日
国内外の別：国内

６．研究組織
(1)研究代表者
延原 肇（NOBUHARA, Hajime）
筑波大学・システム情報系・准教授
研究者番号：80359687