科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２０　日現在

機関番号：　１７１０２
研究種目：　基盤研究(A)
研究期間：　2008～2013
課題番号：　２０２４０００８

研究課題名（和文）ＷＥＢ文書の動的な読解支援システムに関する研究

研究課題名（英文）A Study of User Interface for Document Browsing

研究代表者
石井　久美子（田中久美子）（Kumiko, Tanaka-Ishii）

九州大学・システム情報科学研究科（研究院・教授

研究者番号：１０３２３５２８

交付決定額（研究期間全体）：（直接経費）　　33,100,000円、（間接経費）　　9,930,000円

研究成果の概要（和文）：インターネット文書の読解支援を動的に行うためのユーザインターフェース技術ならびに自然言語処理上の研究を行った。特に、ユーザがweb文書をブラウザで閲覧している際に、文書部分に関してアクションをとると、関連する情報を取得してユーザに動的に提示するweb mash up技術を中心に研究を進めた。内容は、web mash upをブラウザと連携して行うクライアントと、関連情報を抽出するサーバの研究に分かれる。クライアントについては特許を取得し産学連携上の成果につながり、プロトタイプを共に構築した学生が起業しその会社が育った。また、サーバに関しては関連情報を抽出する手法に関して種々の研究成果が挙がった。

研究成果の概要（英文）：In this project, we studied user interface design and natural language processing techniques to build applications to provide related information dynamically to the users when browsing web documents. The study relates to a technique called web mash up which is realized by a client embedded in the browser handling various user actions and a server which communicates with this client and provides the related information to the user. The study of the client was conducted through corporation relations and finally led to a successul entrepreneur company established by a student involved in the project. Moreover, various natural language techniques for extracting related/additional information were studied and this led to academic results including journal and conference papers and invited talks.

研究分野：　自然言語処理、言語インターフェース

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：　ユーザーインターフェイス　自然言語処理　web情報処理　情報抽出

１．研究開始当初の背景

グローバル化の昨今、ユーザは専門性の高い文書や外国語の文書を閲覧しなければならない機会も多く、その読解が簡単でない事も多い。これに対し、ブラウザをより対話的にし、関連情報を動的に表示するなどして、ユーザを支援する事が考えられる。対話的な web ページは、ユーザのローカルなブラウザにプログラムを動的にダウンロードして実行する事を要するため、安全性を考慮する必要があるが､研究開始当初、その枠組みが確立されてきつつあった。この背景を受け、この対話的な機能を利用し、読者を対話的に動的に支援するべく、自然言語処理ならびにユーザインターフェースの観点から研究を行った。

２．研究の目的

研究目的は､対話的な機能の中でも､web mash up について､特に言語処理の観点から可能性を模索する事にあった。web mash up とは、web ページを閲覧中にユーザが文書部分に対してアクションをとると関連情報を示す機能で、異種のデータを組み合わせる(mash up する)事に基づく。web mash up は、ブラウザと連携するクライアントと、クライアントと通信を通して関連情報をユーザに届けるサーバの二つにより実現される。本研究では、クライアントに関しては言語インターフェースの観点から、サーバに関しては自然言語処理の観点から研究を行った。

クライアントについては、ユーザのアクション、また結果の表示の方法など、さまざまな態様が考えられる。また、ブラウザ内でプログラムとして動作するものでもあるため、セキュリティを考慮する必要がある。応用分野がさまざまにある中で、どのような汎用モジュールとして実現するのかを模索する事が課題となる。

サーバについては、自然言語処理の観点から web mash up が特に有用となると考えられる、三つの分野を考える事を通して、可能性を探ることを目標とした。第一は、語学学習に関するもので、たとえば、わからない単語を動的に辞書引きしたり、あるいは文を翻訳したりする。第二は文書の安全性に関するもので、文書の一部が剽窃である事を動的に指摘するなど考えられる。第三は、マーケッティングに関するもので、特に、ユーザが商品など対象物に関する文書を閲覧している時に、関連情報を動的に示す事を目的とした。

３．研究の方法

クライアントに関しては、異なるサーバに対して動作する汎用のソフトウエアを具体的に試作し、ユーザインターフェース上の研究成果を挙げる事を当初は目的とした｡研究期間２年目に、初年度に共に研究した学生が、本研究の成果としてのクライアントのプロトタイプを元に起業したため、この時点で特許や業務上の戦略を重視する事にした。論文に代える形でこの会社からクライアント自体を無償配布する事により、クライアントを広く利用してもらい、フィードバックを得る事を行った。

サーバに関しては、研究開始当初は、三つの分野それぞれにサーバを構築する事を方法として考えていた。しかし、実際に研究を進めてみると、問題とし挙がる要素技術には共通するものがある事がわかってきた。同時に､市場における web mash up の浸透は早く、さまざまな web mash up が個別のサービスとして実現され、研究室で一般的なサーバを立ち上げて運用する事に疑問も生じた。このため、期間半ばに、web mash up に共通する文書処理に必要となる要素技術の研究を行う事を通して研究の観点から本分野に貢献するべく方向を転換した。web mash up では基本的な文書の解析が必要となるが、これに加えて、特に①文書内の変化点の抽出，② 関連情報の抽出のための基礎的な研究が重要となる。

４．研究成果

研究成果としては、
(1)クライアントに関する研究開発
(2)サーバに関する自然言語処理上の研究
(3)本研究を推進する上で、関連して着想に至り実現したソフトウエアなどの研究
の三つが挙げられる。以下それぞれに説明する。

(1)クライアントについては、web mash up に関する具体的なプロトタイプを、初年度に指導学生と共に作成した。クライアントは、携帯端末のブラウザなどでも動作する汎用のもので、基本的には、読解が難しい文書部分をハイライトなどにより指摘すると、サーバが関連情報を送付し、結果をポップアップさせたり、あるいは画面の大きさが小さい場合には文書内に埋め込んだ形で表示する機能を持つ。ハイライトのさせ方や結果の表示の方法で独自の工夫を行った。下図は、携帯電話上で当初のプロトタイプを動作させた web mash up の例である。ユーザは英文の新聞を読む際にわからない部分を指摘し、それを機械翻訳エンジンが翻訳した結果を、ページ内に動的に埋め込んで示している様子が示されている。

クライアントの基本機能に関して、研究期間が始まる直前に特許を申請し、最終年度に取得された。これを元に２年目に当該学生が起業し、３年目にクライアントが無償配布される事となった。会社は順調に育ち、現在大手検索会社に対し、会社の売却に関する交渉が進められている。クライアント自体に関し

ては、このように、産学連携を進めたため、成果の全貌を明らかにする学術論文の成果は、戦略上出す事はしてはいない。しかしながら、本研究の中から、このように会社が育った事は、大きな成果の一つといえる。

(2)サーバに関しては、①文書内の変化点の抽出と、②関連情報の抽出のための文書解析方法の二つが web mash up を行う上で共通する重要な要素技術となる。

①web 上の文書の中には、複数の異なる種の文書が混交する事が多い。日本語の中に英語が混じったり、また、剽窃検知などにおいては著者の異なる文面が混交する。異種文書部分を切り分ける事自体が読解支援につながる他、文書の何らかの解析を行うための前処理としても、切り分けは重要となる。本研究では、文書の種別に応じた切り分けの研究を、機械学習など用いて行った。文書の安全性の観点から剽窃検知、また語学学習の観点から異言語部分の切り分けについて研究し、国内外論文発表など成果を上げた。

②web mash up のサーバは、クライアントから送られてくる情報に対して関連情報を抽出し、送り返すものである。このため、さまざまな関連情報の抽出手法を試みた。語学学習に関しては、文書の難易度判定、漢字検索、辞書逆引きなどの研究を行った。文書の安全性に関しては、主として剽窃検知に取り組んだ。また、マーケッティングに関しては、多言語で同じ内容を表す文書の抽出手法に取り組んだ。そして、これらを支える基礎的な解析技術として、多言語に適用可能な構文解析手法の研究などに取り組んだ。いずれも、国内外での研究発表を行った。

(3)本研究は異なる種の情報を mash up（混交させる）事に関するものであるが、関連して着想した事柄に関して、他の言語インターフェース上の成果が上がった。中でも、コミュニケーション支援ソフトウエアは、複数の成果が上がった。たとえば、ユーザが外国語でコミュニケーションを行う際に、翻訳と平行して pictogram(図や絵)を用いて相手の意思の理解の支援を行う picoTrans など提案した。これは、翻訳サーバと通信しながら、iPad 上で動作するソフトウエアで、言語と図や絵という異種のデータを混交させてユーザ理解を支援する。異種のデータを混交させる着想を web mash up から得ている。本システムに関しては、国際会議で Best Paper Award を受賞する事を初め、多数の成果が上がった。この他にも、いくつかの個別の言語インターフェースに関して関連する成果が上がった。

以上、代表者の異動ならびにライフイベントをまたがる形で、６年の長い研究期間を通して研究を推進した。途中、環境が激変したため、当初の申請案からの方針変更が必要となった。しかし論文発表も順調に行い、本研究に関して学術上招待講演も複数回依頼され、何よりもクライアント技術でベンチャー企業が育った事は大きな成果であると考える。この事から、総じて、本基盤研究に対しての相応の成果が挙がったといえるのではないかと考える。最後になるが、本基盤研究に関わってくださった評価の先生方や事務の皆様に心から御礼申し上げる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１５件）

① Andre Horie and Kumiko Tanaka-Ishii、Sentence hedge detection without cue annotation: A heuristic cue selection approach. 自然言語処理、査読有、21巻、2014、24-40

② Wei Song, Andrew Finch, Kumiko Tanaka-Ishii, Keiji Yasuda, and Eiichiro Sumita 、 picotrans: An intelligent icon-driven interface for cross-lingual communication 、 ACM Transactions on Interactive Intelligent Systems、査読有、3(1)巻、2013 、 1-31 DOI:10.1145/2448116.2448121

③ Andrew Finch, Wei Song, Kumiko Tanaka-Ishii, and Eiichiro Sumita、Speaking louder than words with pictures across languages、AI magazine、査読有、34(2)巻、2013、31-47 DOI: http://dx.doi.org/10.1609/aimag.v34i2.2471

④ Andre Horie, Kumiko Tanaka-Ishii, and Mitsuru Ishizuka、Verb temporality analysis using Reichenbach's tense system、Proceedings of the 24th International Conference on Computational Linguistics (COLING): Posters、査読有、2012、471-482 http://aclweb.org/anthology//C/C12/C12-2047.pdf

⑤ Kumiko Tanaka-Ishii and Hiroshi Terada、Word familiarity and frequency、Studia Linguistica,、査読有、65巻、2011、96-116 DOI: 10.1111/j.1467-9582.2010.01176.

⑥ Kumiko Tanaka-Ishii, Satoshi Tezuka, and Hiroshi Terada、Sorting texts by readability、 Computational Linguistics、 36(2)巻、査読有、2010、203-227 DOI:10.1162/coli.09-036-R2-08-050

⑦ Kumiko Tanaka-Ishii and Julian Godon、Kansuke: A logograph look-up interface based on a few modified stroke prototypes、ACM Transactions on Computer-Human Interaction、16(2)巻、査読有、2009、1-17 DOI:

10.1145/1534903.1534908

⑧ 江原遥 田中久美子、TypeAny: 言語判別を用いた多言語入力システム、自然言語処理、15(5)巻、査読有、2008、151-168
http://repository.dl.itc.u-tokyo.ac.jp/dspace/bitstream/2261/29091/1/v15n5_08.pdf

⑨ Kumiko Tanaka-Ishii and Zhihui Jin、 From phoneme to morpheme: Another veri_cation in English and Chinese using corpora、Studia Linguistica、62(2)巻、査読有、2008、224-248
DOI: 10.1111/j.1467-9582.2007.00138.x

〔学会発表〕(計３９件)

① 中谷洸樹, Andrew Finch, 田中久美子, and 隅田英一郎、確率的ブロック編集距離、言語処理学会大会、2014.3.18、北海道

② 村脇有吾, 粟飯原俊介, 原田泰佑, 長尾真, and 田中久美子、意味的逆引き辞書『真言』におけるスコア付け、 言語処理学会大会、2014.3.18、北海道

③ 田中久美子、User Interface Design for International Communication 、International Collaboration Symposium (hosted by Prof. Yves Lepage, Waseda University) 、2013.11.11、北九州、基調講演

④ 粟飯原俊介, 長尾真, and 田中久美子、意味的逆引き辞書『真言』、言語処理学会大会、2013.3.14、名古屋

⑤ Andre Horie, Kumiko Tanaka-Ishii, and Mitsuru Ishizuka 、Verb temporality analysis using Reichenbach's tense system、 24th International Conference on Computational Linguistics (COLING)、2012.12.14、Mumbai

⑥ Hiroshi Yamaguchi and Kumiko Tanaka-Ishii 、Text segmentation by language using minimum description length 、50th Annual Meeting of the Association for Computational Linguistics (ACL)、2012.7.9、Jeju

⑦ Kotaro Kitagawa and Kumiko Tanaka-Ishii、Relational lasso: An improved method using the relations among features、5th International Joint Conference on Natural Language Processing (IJCNLP)、2011.11.9、Chiang Mai

⑧ Andrew Finch, Wei Song, Kumiko Tanaka-Ishii, and Eiichiro Sumita 、Source language generation from pictures for machine translation on mobile devices、8th International Workshop on Natural Language Processing and Cognitive Science (NLPCS)、2011.8.20、Copenhagen

⑨ Andrew Finch, Wei Song, Kumiko Tanaka-Ishii, and Eiichiro Sumita 、picoTrans: Using pictures as input for machine translation on mobile devices、22nd International Joint Conference on Artifcial Intelligence (IJCAI), 2011.7.16、Barcelona

⑩ Wei Song, Andrew Finch, 田中久美子, and 隅田英一郎、 picotrans: A user interface using content word-based source sentence generation for machine translation 、言語処理学会大会、2011.3.10、豊橋

⑪ 北川浩太郎,田中久美子、木構造に基づく係り受け解析、 言語処理学会大会、2011.3.10、豊橋

⑫ 溝江将,田中久美子、単独剽窃検知-所与の一文書だけから剽窃箇所を推定する-、言語処理学会大会、2011.3.9、豊橋

⑬ Wei Song, Andrew Finch, Kumiko Tanaka-Ishii, and Eiichiro Sumita、picoTrans: An icon-driven user interface for machine translation on mobile devices、 16th International Conference on Intelligent User Interfaces (IUI)、2011.2. 13、Palo Alto

⑭ 田中久美子、Two Language Learning Applications Using Search Engines、Invited Talk at Google、2010.9.30

⑮ Kotaro Kitagawa and Kumiko Tanaka-Ishii 、Tree-based deterministic dependency parsing: An application to Nivre's method、48th Annual Meeting of the Association for Computational Linguistics (ACL) 、2010.7. 13、Uppsala

⑯ 北川浩太郎,田中久美子、段階的な部分木間の構造判定に基づく決定的係り受け解析、言語処理学会大会論文集、2010.3.6、東京

⑰ 田中久美子、言語処理を用いた語学教育支援-二つの観点からの取り組み-、2009年度科研・合同シンポジウム、2009.10.15、東京

⑱ Dani Yogatama and Kumiko Tanaka-Ishii、Multilingual spectral clustering using document similarity propagation、International Conference on Empirical Methods for Natural Language Processing (EMNLP) 、 2009.8.6 、Singapore

⑲ 手塚智史,寺田博視,田中久美子、相対的観点に基づく類似難易度文書検索システムの構築、言語処理学会大会、2009.3.3、鳥取

⑳ 寺田博視,田中久美子、文書の難易順序判定法、言語処理学会大会ワークショップ「教育・学習を支援する言語処理」、2008.3.21、東京

㉑ 周安平,田中久美子、Omniget：第三者情報を提示するブラウザ内ブラウザ、 言語処理学会大会、2008.3.20、東京

㉒ 寺田博視,田中久美子、 単語の親密度と単語頻度の関係に関する一考察、言語処理学会大会、2008.3.19、東京
㉓ Kumiko Tanaka-Ishii、Predictive Entry Systems using Statistical Language Models 、Japan-America Frontiers Engineering Symposium, funded by JST and PNAS、2008.11.19、神戸
㉔ Yo Ehara and Kumiko Tanaka-Ishii、Multilingual text entry using automatic language detection、3rd International Joint Conference on Natural Language Processing (IJCNLP)、2008.1.7、Hyderabad

〔図書〕（計　４件）
1.田中久美子、長谷川寿一監修、東京大学出版会、言語の分節に普遍的に観察される統計的性質: 音素から形態素へ、単語へ、そして句へ．こころと言葉、2008、 211-226

〔産業財産権〕
○出願状況（計　３件）

名称：発話の評価方法及び装置、発話を評価するためのコンピュータプログラム
発明者：Daniel Heffernan、田中久美子
権利者：同上
種類：特許
番号：特願 2013-201242
出願年月日：２５年９月２７日
国内外の別： 国内

名称：検索システム
発明者：石井久美子、長尾真、粟飯原俊介
権利者：同上
種類：特許
番号：特願 2013-142386
出願年月日：２５年７月８日
国内外の別： 国内

名称：Check System, Information Providing System, and Computer-Readable Information Recording Medium Containing a Program,
発明者：石井久美子
権利者：同上
種類：特許
番号：PCT/JP2008/061729
出願年月日：２０年６月２７日
国内外の別： 国外

○取得状況（計　４件）

名称：ウェブ文書を表示する端末装置が実行するプログラム
発明者：石井久美子
権利者：同上
種類：特許
番号：特許第 5360842 号
取得年月日：２５年９ 月１３ 日
国内外の別：国内

名称：確認システム、情報提供システム、ならびに、プログラム
発明者：石井久美子
権利者：同上
種類：特許
番号：特許第 4877831 号
取得年月日：２３年１２ 月９日
国内外の別：国内

名称：文字入力装置、文字入力方法、ならびに、プログラム
発明者：石井久美子
権利者：同上
種類：特許
番号：特許第 4706021 号
取得年月日：２３年３月２５日
国内外の別：国内

名称：漢字仮名交じり入力装置、漢字仮名交じり入力方法、ならびに、情報記憶媒体
発明者：石井久美子
権利者：同上
種類：特許
番号：特許第 4423369 号
取得年月日：２１年１２月１８日
国内外の別：国内

〔その他〕
ホームページ等
http://www.cl.ait.kyushu-u.ac.jp/Research.html

本プロジェクトから育ったベンチャー企業の HP
http://www.popin.cc

６．研究組織
(1)研究代表者
石井　久美子　（ISHII, kumiko）
（田中　久美子）（TANAKA, kumiko）
九州大学・システム情報科学研究科（研究院）・教授
研究者番号：１０３２３５２８

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：