科学研究費助成事業　　研究成果報告書

平成　２６　年　　４　月　２８　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１１６

研究課題名（和文）集合知とウェブ知識の有機的循環化技術の開発

研究課題名（英文）Development of technologies for organic cycle of collective intelligence and knowledge on the Web

研究代表者
浅野　泰仁（Asano, Yasuhito）

京都大学・情報学研究科・准教授

研究者番号：２０３６１１５７

交付決定額（研究期間全体）：（直接経費）　　3,300,000円、（間接経費）　　990,000円

研究成果の概要（和文）：本研究では，集合知をさらなる発展段階に導くための基盤技術として，(1) 集合知の情報構造を利用してウェブから集合知を補完する手法と，(2) 得られた知識を整理して提示する手法を構築することを目指した．成果は，(1)としては，Wikipediaに不足しているテキスト情報をウェブから取得する手法，Wikipediaに不足している画像情報、特にエンティティ間の関係を説明するものをウェブから取得する方法，などの提案であり，(2)としてはウェブから取得した、エンティティ間の関係を説明するような画像をWikipediaの知識と合わせて提示する方法などの提案である．

研究成果の概要（英文）：In this work, we have aimed to construct the following two fundamental technologies (1)-(2) for developing collective intelligence: (1) methods for complementing collective intelligence from the Web utilizing information structure of the collective intelligence, and (2) methods for organizing and presenting obtained knowledge. The results for (1) include a method for complementing lacked text information on the Wikipedia from the Web and a method for finding lacked visual information on the Wikipedia, especially images for explaining the relationship of specified two entities, from the Web. The results for (2) include a method for presenting images collected from the Web for explaining the relationship of specified two entities, by combining knowledge on Wikipedia for the relationship.

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：　知識発見　情報検索　知識体系化　集合知　ウェブ

１．研究開始当初の背景

ウェブは情報検索および知識の獲得のために欠かせないものである．特に，近年では Wikipedia, Yahoo!知恵袋, 価格.com 等に代表される集合知サービスが，多くの人々にとって知識獲得や意志決定に重大な役割を果たすようになってきている．また，これらの集合知サービスでは，ある程度情報が構造化されているので，この構造を用いてセマンティック・データを抽出する研究も数多く行われてきた．ただし，抽出できる知識は非常に限られている．

これらの集合知の構築は人手に依存しており，現在の集合知に含まれている知識を調査しつつ，新しい知識をその集合知の構造に合わせて整理して加えていく必要があるので，数多くのユーザの多大な労力なしには成り立たない．このため，一般のウェブページやブログなどに比べて，集合知に含まれる知識は限定的なものにならざるを得ない．また，多くの集合知サービスは誰でも編集に参加できるため，しばしばコンテンツの信憑性や可読性が低くなる傾向にある．

申請者は，これまでウェブ情報発見分野や，Wikipedia から関係に関する知識を体系化する研究について数多くの成果があり，NICT 委託研究の「信憑性プロジェクト」では「Wikipedia の画像信憑性問題」に取り組んで成果を上げている．従って，ウェブや集合知サービスから自動的に情報を抽出する技法や，それらを整理・体系化する技法に詳しい．特に，集合知の情報構造を利用して，ウェブから「関係」に関する知識を検索する手法の開発に着手していた．

２．研究の目的

本研究では，集合知をさらなる発展段階に導くための基盤技術として，集合知の情報構造を利用してウェブから集合知を補完する知識を検索する手法と，得られた知識を整理して提示するとともに，先程の補完知識検索に役立てる手法を構築することを目指した．

これによって，補完知識検索手法の性能がさらに向上し，結果として補完知識も増えるという循環ができる．すなわち，集合知と一般のウェブページに含まれる知識が有機的に循環していく基礎が形成され，集合知の発展のみならず，ユーザに対してそのコンテンツの信憑性を判定する支援や，同じ内容でもより可読性の高いコンテンツの推奨が可能となると考えられる．さらに集合知自体にとどまらず，一般のウェブページのコンテンツ発展にも寄与していくことが期待される．

３．研究の方法

Wikipedia を含む特定の集合知を対象として，「集合知を補完する知識の検索手法」と，「補完知識の整理・提示手法」の確立を，目指した．

(1)集合知を補完する知識の検索手法

集合知を補完する知識を検索するには，その集合知サービスにおける情報の構造(Wikipedia ではカテゴリの構造や，各項目のセクション構造，各項目間のリンク構造など)と，各コンテンツの言語的構造を統合したモデルを構築し，このモデルに基づいて集合地上の知識をウェブ上のコンテンツの知識と比較しつつ，必要な情報を検索する手法が必要と考えられる．本研究では，申請者がこれまでウェブ情報発見分野および Wikipedia 関連研究で挙げてきた研究成果を応用することで，この手法の提案及び確立を目指した．

(2) 補完知識の整理・提示手法

申請者が NICT 委託研究「信憑性プロジェクト」で得られた知見と手法，そして申請者が所属する京都大学吉川研究室で育まれてきた知識の体系化技術を応用して，その「補完知識」を整理する．具体的には，もとの集合知に追加すべき知識や，もとの集合知と矛盾するような知識，あるいはもとの集合知に存在する知識と同じことについて述べているがよりわかりやすい図解や記述などを抽出し整理・提示する手法を確立することを目指した．

４．研究成果

(1)集合知を補完する知識の検索手法

(1-a) Wikipedia に不足しているテキスト情報をウェブから取得する方法

Wikipedia の更新は人手に頼っているため、Wikipedia には最新の情報や重要な情報が不足していることが往々にして存在する。本研究では、与えられた Wikipedia 記事に不足しているテキスト情報をウェブから取得する手法を構築した．具体的には，まず，Wikipedia の記事と関連ウェブページ集合それぞれに LDA によるトピック解析を行うことで両者のトピックの差を解析し，Wikipedia に存在しない，または情報が不足しているトピックを検出する．次に，このトピック解析の情報を元に，関連ウェブページ集合からそのようなトピックに最も適合した文章を抽出する．これによって，Wikipedia のスタブ(作りかけの)記事に対しても，ウェブの情報を用いて内容を補完することが可能になった．(雑誌論文 5，学会発表 6)

(1-b) Wikipedia に不足している画像情報、特にエンティティ間の関係を説明するものをウェブから取得する方法

Wikipedia の記事には著作権上の問題から画像が少ないことがよくある．特に，「日本と石油の関係を説明する画像」のように，エンティティ間の関係を説明する画像はまれである．本研究では，代表者らが提案した，Wikipedia の記事間のリンク構造のようなエンティティ間のネットワークにおける関係の強さを測るための「減衰流モデル」で求められる「関係を成り立たせているエンティテ

ィ」の情報を用いて，ウェブからエンティティ間の関係を説明する画像を取得する方法を提案した．(雑誌論文 2,4,7，学会発表 9)

本手法の理論的背景となっているのは減衰流(generalized flow)だが、その基礎となっているネットワーク理論に関する(ネビュラ賞を受賞した Kleinberg らが著した)図書(1)を和訳した。さらに、減衰流計算の高速化の手法についても研究し、貪欲法と適切な初期フローを選択するヒューリスティクスによって高速かつ高精度でこれを計算できる近似手法を構築した．(学会発表 4,14)

(1-c) Wikipedia に不足している、クラシック音楽の内容記述をウェブから取得する方法

クラシック音楽を鑑賞する際には，その曲の音楽的な内容や構造に関する記述を読むことで曲に対する理解を深めることができる．Wikipedia にはごく有名な一部の曲に関しては内容記述が充実しているものもあるが，多くの曲については内容記述がほとんどない．また，内容記述を含むウェブページは存在するが，単に曲名や、それに加えて「内容」「構造」といったキーワードおよび楽器名を用いたとしても，一般のウェブ検索では効率的に見つけることが困難である．本研究では，クラシック音楽の楽曲に関するウェブページを 8 種類に分類するラベル付けを提案し、それを利用した Labeled LDA を用いることでクラシック音楽の内容記述をウェブから収集する手法を提案した．(学会発表 1,11,12)

(1-d) 複数のブログ上で盛り上がった話題のグラフマイニング解析による役割知識補完

ある話題に関するブログ記事を集積したものは，その話題に関する集合知を形成している．本研究では，重要度が高いのに既存のマスコミにはあまり取り上げられることのないような，複数のブログ上でのみ盛り上がった話題を解析した．これらの話題は複雑にリンクし合ったブログサイト上で時系列を追うごとに広範囲に伝播していくが，このような構造に汎用的に適用できる「時間グラフパターンマイニング」のフレームワークを提案し，これを機械学習に適用した分析によって，話題の提示役・盛り上げ役・まとめ役といったブログサイトの役割を自動的に分類することが可能になった．この役割を見ることによって，話題に関する集合知であるブログの個々の記事からは得られない，その形成に関する知識を補完することができるようになる．さらに，ここで得られた時間グラフパターンマイニングは，Amazon を初めとする推薦グラフの分析にも有用であることを示した。(雑誌論文 1)

(1-e) 画像投稿サイトのタグ補完技術を用いた、類似画像検索手法

Flickr などの画像投稿サイトでは，画像にタグを付けることができる．これらの投稿画像とタグは，画像に関する集合知と見なすことができる．このタグを利用したキーワード画像検索はすでに一般的だが，画像を入力とする類似画像検索には十分に利用されているとは言えなかった．原因としては，タグが投稿者本人によって付与されるため，不十分な場合や不適切な場合が多く存在することが上げられる．本研究では，まずタグを与えていない入力画像に対して，それらの画像と画像的に類似した画像の集合を求め，それらに付与されたタグ集合と画像集合の関係を表すグラフを分析することによって，入力画像にふさわしいタグ情報および画像集合中のタグの適切さに関する情報を補完し，入力画像と意味的に類似する画像のランキングを得る手法を提案した．
(雑誌論文 3,6,8，学会発表 2,5,7,10)

(1-f) レビューサイトの特異なレビュアーの発見および早期レビューからの将来予測手法

レビューサイトは本などの商品やレストランなどのサービスを購入する際に大きな影響力を持つに至っている集合知である．しかし，ステマや誹謗中傷を行う悪意のあるレビュアーや，マニアックすぎて安価な商品や一般のユーザーを見下すレビュアーも存在する．これらの特異なレビュアーの存在は，特にレビューが少ない発売直後の商品や開店間もない店舗に大きな影響を及ぼしてしまう．特異なレビュアーの情報を補完することで，レビューサイトを多くの人々がより安心して活用することができるようになるはずである．本研究では，レビュアーと対象のなす 2 部グラフを分析する新しい手法として，特異なレビュアーの発見を行い，同時にその影響を弱めることで一般のユーザーの意識に近いレビュー集約結果を得る手法を提案した．また本手法によって，レビューサイト自体の運用期間が長ければ，過去のデータを使うことで，早期レビューに対しても特異なレビュアーの影響を検出し，将来一般ユーザーが多くなったときのレビュー集約結果を予測することができることを示した．(学会発表 8)

(2) 補完知識の整理・提示手法

(2-a) ウェブから取得した、エンティティ間の関係を説明するような画像を Wikipedia の知識と合わせて提示する方法

(1-b)で提案した，「Wikipedia に不足しているエンティティ間の関係を説明するものをウェブから取得する方法」によって得られた画像集合を，代表者らが以前提案した関係理解支援システムと組み合わせて提示するシステムを構築した．上記の関係理解支援システムは，二つのエンティティ間の関係を，

その関係を成り立たせる重要なエンティティのパスの集合で表し，提示する．例えば「日本」と「石油」を二つのエンティティとすると，パスの例は「日本」→「秋田県」→「油田」→「石油」などである．このパスが存在する理由は，秋田県には日本でもまれな油田が存在することに起因していると考えられるが，その場合，提案システムはこのパスに関連する画像をパスに結びつけて表示することができる．また，パス集合すなわち関係全体に関連する画像を提示することもできる．これによって，システム利用者は関係の強さとその理由を推測しやすくなる．(雑誌論文 2)

(2-b) 画像投稿サイトのタグを用いた、類似画像ランキングの提示とユーザーフィードバックによる改善手法

(1-e)で提案した，入力画像と意味的に類似した画像を検索する手法に基づいて，これらの画像のランキングを提示し，ユーザーが正負のフィードバックを与えることを可能にするシステムを提案した．さらに，このフィードバックを(1-e)の手法の一部に組み込み，ランキングを改善する手法を提案した．(雑誌論文 8，学会発表 2,5,10)

(2-c) 歴史の補完的知識発見を目指した Wikipedia 上の歴史事象の体系化

「歴史に学ぶ」とは，未来のために過去の歴史事象知識を活かすことである．Wikipedia は多くの歴史事象に関する集合知と見なすこともできるが，これを利用して歴史に学ぶためには，本質的に類似した事象の発見や，その相違点の検出が必要になる．これを人手で実現するには，非常に多くのページを閲覧する必要があり，現実的ではない．また自動的に行うにしても，類似した単語の出現のみに頼る方法では，字面は似ているが本質的に異なる事象や，本質的に類似しているのに時代背景の差異などによって用語が異なる事象などは発見できない．本研究では，事象のそれ以上分割できない粒度の記述を極小事象とし，一般の事象を複数の極小事象とその関係を表すグラフ構造からなる複合事象とみなし，さらにこのグラフ構造を複数の事象間で比較することで，単語のみならず事象の本質的順序や構造を考慮した類似事象の発見を目指した．今回は，上記の極小事象・複合事象の具体的なモデルと，Wikipedia 記事から極小事象集合とそれらの間の関係を，談話構造解析手法の応用によって抽出する手法を提案した．(学会発表 3)

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 8 件）

1. Yasuhito Asano, Taihei Oshino, Masatoshi Yoshikawa. “Time Graph Pattern Mining for Network Analysis and Information Retrieval”. IEICE Trans. Inf. Syst., Vol E97-D, No.4., pp.733-742, 2014.

2. Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. “Mining Knowledge on Relationships between Objects from the Web”. IEICE Trans. Inf. Syst., Vol E97-D, No. 1, pp.77-88, 2014.

3. Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa. “Tag Quality Improvement for Social Image Hosting Website”. IPSJ Transactions on Databases, Vol. 6, No. 3, pp.177-186, 2013.

4. Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. “A Generalized Flow-Based Method for Analysis of Implicit Relationships on Wikipedia”. IEEE Trans. Knowl. Data Eng. 25(2): 246-259 (2013).

5. Damien Eklou, Yasuhito Asano, Masatoshi Yoshikawa. “How can the Web help Wikipedia? A Study of Information Complementation of Wikipedia by the Web”. IPSJ Transactions on Databases, Vol. 5, No. 3, pp. 64-72, 2012.

6. Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa. “Improving Content-based Social Image Retrieval Based on an Image-tag Relationship Model”. IPSJ Transactions on Databases, Vol. 5, No. 3, pp. 117-125, 2012.

7. Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. “Mining and Explaining Relationships in Wikipedia”. IEICE Trans. Inf. Syst., Vol E95-D, No. 7, pp. 1918-1931, 2012.

8. Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa. “Re-ranking Content Based Social Image Search Results by Multi Modal Relevance Feedback”. DBSJ Journal, Vol. 11, No. 1, pp. 67-72, 2012. 5.

〔学会発表〕（計 13 件）

以下の 9 件は査読有

1. Taku Kuribayashi, Yasuhito Asano, Masatoshi Yoshikawa. “Ranking Method Specialized for Content Descriptions of Classical Music”. In Poster Proc. of WWW 2013, pp.141-142, 2013.

2. Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa. “Potential Semantics in Multi-modal Relevance

Feedback Information for Image Retrieval”. In Proc. of the 37th Annual IEEE Computer Software and Applications Conference (COMPSAC 2013), pp.830-831, 2013.

3. Minoru Naito, Yasuhito Asano, Masatoshi Yoshikawa. “A Graph Model of Events Focusing on Granularity and Relations Towards Organization of Collective Intelligence on History”. In Proc. of WEB 2013, pp. 111-115, 2013.

4. Yusuke Nojima, Yasuhito Asano, Masatoshi Yoshikawa. “Greedy Approximation Algorithms for Generalized Maximum Flow Problem towards Relation Extraction in Information Networks”. In Proc. TJJCCGG 2012, 2012.

5. Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa. “A Re-ranking by Multi-Modal Relevance Feedback for Content-Based Social Image Retrieval”. In Proc. APWeb 2012, pp. 399-410, 2012.

6. Damien Eklou, Yasuhito Asano, Masatoshi Yoshikawa. “How the Web can help Wikipedia: A Study of Information Complementation of Wikipedia by the Web”. In Proc. ICUIMC 2012, 9, 2012.

7. Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa. “Ranking Content-Based Social Images Search Results with Social Tags”. In Proc. AIRS 2011, pp.147-156, 2011.

8. Kazuki Tawaramoto, Junpei Kawamoto, Yasuhito Asano, Masatoshi Yoshikawa. “A Bipartite Graph Model and Mutually Reinforcing Analysis for Review Sites”. In Proc. DEXA 2011, pp.341-348, 2011.

9. Xinpeng Zhang, Yasuhito Asano, Masatoshi Yoshikawa. “Towards Improving Wikipedia as an Image-rich Encyclopedia through Analyzing Appropriateness of Images for an Article”. In Proc. APWeb 2011, pp. 200-212, 2011.

以下の 5 件は査読なし

10. Jiyi Li, Qiang Ma, Yasuhito Asano, Masatoshi Yoshikawa. “Multi Modal Relevance Feedback for Content Based Social Image Retrieval”. The 5th International Workshop with Mentors on Databases, Web and Information Management for Young Researchers (iDB WorkShop 2013), No.15, 2013.

11. 栗林拓, 浅野泰仁, 吉川正俊. “クラシック音楽の内容記述のウェブからの収集手法”. 第6回データ工学と情報マネジメントに関するフォーラム(DEIM 2014), 2014.

12. 栗林拓, 浅野泰仁, 吉川正俊. “クラシック音楽の内容記述に特化した検索手法”. 第5回データ工学と情報マネジメントに関するフォーラム(DEIM 2013), 2013.

13. 内藤 稔, 浅野 泰仁, 吉川 正俊. “エンティティ間の類似関係取得のための Wikipedia 事象モデル構築手法に関する考察”. 第4回データ工学と情報マネジメントに関するフォーラム(DEIM 2012), 2012.

14. 野島 裕輔, 浅野 泰仁, 吉川 正俊. “情報ネットワークにおける関係の抽出のための減衰流の計算の高速化”. 第4回データ工学と情報マネジメントに関するフォーラム(DEIM 2012) , 2012.

〔図書〕（計 1 件）

(1) 浅野孝夫、浅野泰仁(訳). 「ネットワーク・大衆・マーケット -現代社会の複雑な連結性についての推論-」共立出版 (原書: D. Easley, J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge University Press) , 2013 年 6 月.

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕

ホームページ等

Enishi のページ

６．研究組織

(1)研究代表者

浅野　泰仁（ASANO, Yasuhito）
京都大学・大学院情報学研究科・特定准教授
研究者番号：20361157

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：