1F2-4

# Web 文書に対するマーキングからの個人知識の獲得

Supporting Taxonomic Knowledge from Personal Marked Expressions on the Web Pages

松岡 有希[*1*2*3] 坂本 竜基[*1] 伊藤 禎宣[*1*4] 武田 英明[*1*2*3] 小暮 潔[*1]
Yuki Matsuoka Ryuuki Sakamoto Sadanori Ito Hideaki Takeda Kiyoshi Kogure

[*1] 株式会社国際電気通信基礎技術研究所 Advanced Telecommunications Research Institute International
[*2] 総合研究大学院大学 The Graduate University for Advanced Studies
[*3] 国立情報学研究所 National Institute of Informatics
[*4] 東京農工大学 Tokyo University of Agriculture and Technology

In this paper, we proposed the system that classified the web page based on personal preference. The system provides web page links related to the selected sentence on the web page by user. If, user selects a web page link from recommended links, the system defines both selected sentence and a web page link as personal preference, and makes a personal knowledge using it. As a result, personal marked expressions on the web pages are used by folksonomy's tag. So, the knowledge sharing is achieved among users.

## 1. まえがき

近年 Web コンテンツを分類する手法としてフォークソノミーが提唱され，様々なソーシャルブックマークサービス(del.icio.us, Furl, はてなブックマークなど)が登場している．これらのサービスで，ユーザは「タグ」と呼ばれるキーワードを用いて Web コンテンツを分類し，他のユーザとタグを共有する．タグはコンテンツに対するメタデータであり,タグを使って今までとは異なる方法で情報のブラウジングや検索ができるようになった．しかし,他のユーザによって付与された多様なタグは同義語や同音異義語の問題を抱えていることから，ユーザが情報を探す際，役に立たない，あるいは適切でないメタデータになってしまう場合がある．しかし，従来の複雑な階層分類やカテゴライゼーションスキーマを作成するコストに比べて，ユーザがタグを付与するコストのほうがはるかに少なくてすむので，タグによる Web コンテンツの分類は有益とされている．[Mathes 2004].

一方，坂本は第 19 回人工知能学会全国大会で，タグの手入力を簡略化するために Web ページへのマーキングをタグと定義した“イロノミー”を提案した［坂本 2006］.マーキングに関しては，Web ページの文書をマウスでなぞり読みしたテキストはユーザが興味を持つキーワードを多く含んでいる，との報告[土方 2002]があることから，イロノミーを使用すれば，ユーザの興味に基づいた Web ページの分類が期待できる．

本論文では，イロノミーを運用した結果から得られた知見を基に，ユーザの個人知識に基づいた Web ページ分類の実現に向けてシステムを改良する．以下，2 章にて個人知識について解説し，3 章にてイロノミーの概要および分析結果について述べ，4 章にて新しいマーキングシステムについて述べ，5 章でまとめを行う．

## 2. 個人知識

[鷹城 2002]において，それぞれの人間の経験や興味･関心によって知識の想起のされ方が異なることから，このような個々人によって異なる概念間の連想的想起の繋がり方やその強さの構造のことを個人知識と定義している．例えば，「エージェント」というキーワードを人に与えたとき，「マルチエージェント」のことを思い浮かべる人もいれば，「協調学習」という学習法のことを思い浮かべる人もいるだろう．こうした個人独自の語の想起，すなわち個人知識に従って Web ページを分類すれば，人間の思考に沿った情報の整理ができるようになる，と考える．そこで本論文では Web ページに対するインタラクティブな操作からユーザの個人知識を構築するシステムを提案し，ユーザの個人知識に基づいた Web ページの分類を行うことを目標とする．

連絡先：松岡有希，〒101-8430 東京都千代田区一ツ橋 2 丁目 1 番地 2 号 m-yuki@grad.nii.ac.jp

## 3. イロノミー

本章では，第 19 回人工知能学会全国大会における Web 大会支援システムの一機能として開発･運用したシステム“イロノミー”について述べる．イロノミーはユーザが Web ページ内の重要と思われる箇所にマーキングするシステムであり，マーキングにどのような特長があったのかについて調べた．

### 3.1 イロノミーの概要

イロノミーは明治大学の斉藤孝教授が提唱した三色ボールペン法に基づいて，Web ページ内の文字列をマウスカーソルで選択することでマーキングができるシステムである．三色ボールペン法とは,客観的および主観的に重要な箇所に色付きのアンダーラインを引く読書法である．ユーザは自身が重要だと思う箇所にマーキングを付けると，そのデータは他のユーザと共有される．イロノミーではマーキングがフォークソノミーのタグの役割を果たしてり，自身が付けたマーキングだけでなく，他人が付けたマーキングの情報を検索したり，閲覧したりすることできる．

### 3.2 マーキングされた文字列の特徴について

イロノミーの運用で得られたデータを基に分析をした結果，マーキングされた文字列（全ユーザのマーキングデータを使用した）を調べたところ，tfidf [Salton 1991]値の高い語が多く含まれていることが分かった[松岡 2006]．よって,ユーザは三色ボールペン法に従って，文書内で特徴語とされるものにマーキング

していることが分かり，マーキングがユーザにとって特徴語を示す手法として適していることが分かった．

本節ではマーキングされた文字列の中で tfidf 値が低い単語に着目する．表 1 の左列はマーキングされた文字列に含まれる単語の中で，tfidf 値が低い順に上位 10 個を並べたものである．一般的に tfidf は文書内の特徴語を抽出する際に用いられるため，タグの自動抽出をした場合，tfidf 値が低い語は無視される可能性が高い．そこで,tfidf 値が低い単語はタグとして有益でないのかどうかを調べるため，既存のソーシャルブックマークサービスである，はてなブックマークのタグと比較した（表1）．その結果，「研究」，「情報」，「人間」の 3 単語は，既存のタグにも存在した．よって，マーキングされた文字列から得られる単語は tfidf 値が低くてもタグの役割を果たすことが分かる．このことから，ユーザが自らタグを記述しなくても，マーキングで代用できることが分かった．

**表 1　マーキングされた文字列に含まれる単語（tfidf 値が低い単語）とはてなブックマークのタグが同じかどうかの判定**

| | |
|---|---|
| 研究 | ○ |
| 手法 | × |
| パターン | × |
| 適用 | × |
| 情報 | ○ |
| 特徴 | × |
| 複数 | × |
| 方法 | × |
| 個人 | × |
| 人間 | ○ |

一方，表 1 において，マーキングされた文字列に含まれる単語で，はてなブックマークのタグとして存在していなかったタグを見ると，Web ページを分類するにあたりふさわしくない単語がある．例えば「複数」の場合，もとのマーキングされた文字列は「複数のセンサ」である．この場合，「センサ」という語のほうがタグとしては好ましい．また,「手法」の場合，マーキングされた文字列は「利用者一人が提示できる順序の数が一つだけという制限を取り除いた手法」である．この文字列は文章内で重要な箇所を示しているが，長い文字列であるため，ページ分類用のタグとしてどの単語に注目すべきなのかはマーキングをしたユーザのみが判断できる．

上記の考察より，イロノミーではユーザがマーキングをする際，三色ボールペン法を用いるように指示したため，文書の内容として重要な箇所が選ばれたものの，タグとして好ましくない単語が含まれる，あるいはマーキングされた文字列が長くなれば，どの単語にユーザが注目しているのかが分からないという問題が生じることが分かった．

## 4. システムの設計と動作

イロノミーではユーザが Web ページにマーキングを付与したり，その情報を共有するというフォークソノミーの機能を実現していた．運用結果を分析したところ，マーキングされた文字列は文書内の特徴語が多く含んでいたことから，マーキングの有効性が示すことができた．しかし,マーキングされた文字列からはタグとして好ましくない単語が含まれている，あるいは選択された文字列が長くなるほど，どの単語にユーザが着目しているのかが分からないといった問題があった（3.2 節）．よって，マーキングをする際，ユーザがどの部分に注目しているのかが分かるようなシステム設計が必要である．

次に，ユーザの知識を獲得するためのシステム設計について述べる．野中は著書で，「知識が人間の行為と本質的に関係している」と言及している[野中　1996]．つまり,人間は何らかの行為をした結果，それに伴う情報は人間の知識と言えるということを意味している．これに従って人間が知識を獲得するプロセスをシステムの設計に取り入れる．システムは，まずユーザに Web ページ内で興味を持った箇所を選択してもらう．次にユーザが選択した箇所に対してそれに関連するページを一覧表示し，この中から興味のあるページを選択してもらう．システムは，ユーザが興味を持った箇所と選択した関連ページを，行為に伴う情報と判断し，これらを知識と定義する．システムが提供する動作は以下のようになる．

- 情報の表示フェーズ
  - ユーザが Web ページ内の文字列をマウスカーソルで選択すると，選択文字列に関する Web ページのリンクの一覧をポップアップ表示する（図 1）
- ユーザ知識の獲得フェーズ
  - 表示された Web ページのリンク一覧からユーザが興味のあるページを選び，リンクをクリックしてページ遷移した時点で，システムはユーザが選択した文字列をマーキングする（図 2）

システムが提供する，情報の表示および知識獲得の機能をユーザが繰り返し使用することで，システムはユーザの知識を獲得し，個人知識を構築していく．さらに,3.2 節より，長い文字列にマーキングされると，ユーザが文字列内のどの部分に興味を持っているのかが分からなくなるという問題が生じることから，情報の表示部分（図 1）で選択する文字列の長さを長くすればするほど，表示する情報を絞ることで，長い文字列を選択しにくくする．具体的には，選択文字列に含まれる単語のアンド検索を行う．選択文字列に含まれる単語が多くなればなるほど一致する関連ページがなくなっていくので長い文字列は選択できなくなると思われる．さらにこの方法を使えば，選択文字列に含まれる単語の共起関係からユーザがよく使う単語が分かるようになり，タグとして好ましい単語を把握できるようになる．

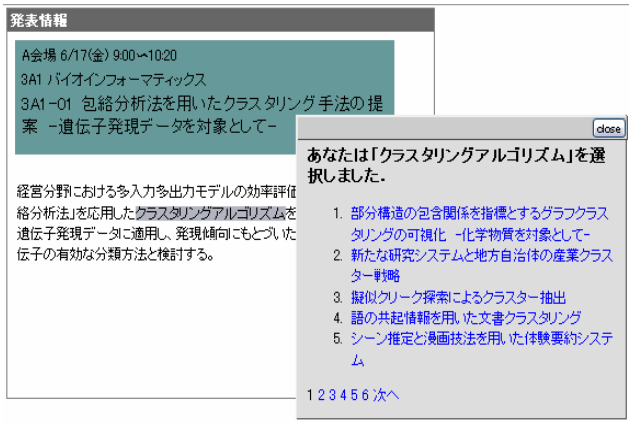


**図 1　情報の表示**

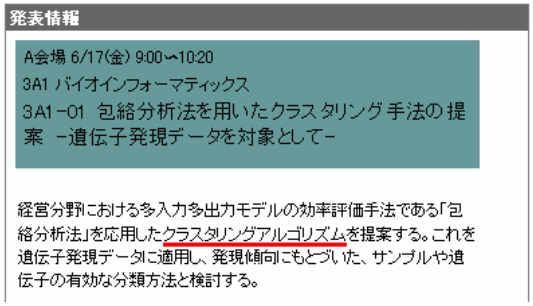


**図 2　ユーザ知識の獲得**

一方，システムによってマーキングされた文字列は，フォークソノミーのタグの役割を果たす．ユーザに対してタグに関する情報は以下のように提供する．

- タグに関する情報の表示フェーズ
  - ユーザがマウスカーソルでマーキングの上をなぞると，マーキングされた文字列が他のページのマーキング文字列に含まれている場合，そのページへのリンクとマーキングの一覧をポップアップ表示する（図 3）．

システムによって付けられたタグはユーザの知識を表しており，ユーザはタグに関連する自身および他人の知識を知ることができるようになる．

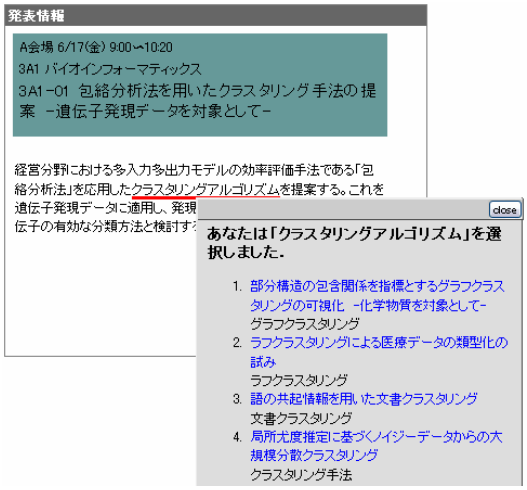


**図 3　タグに関する情報表示**

このように本システムは，情報の表示フェーズおよびユーザ知識の獲得フェーズ，タグに関する情報の表示フェーズの3つの機能を持つ．ユーザは Web ページに対して興味のある箇所を選択し，システムが表示する情報の中から気に入った情報を選択する，という作業を繰り返すことで，知らないうちに個人知識が構築されていく．システムはユーザの知識を獲得することで個人知識を構築し，それを用いてユーザに適した情報推薦を行う．また,ユーザの知識は Web ページにおいてマーキングとして表示され，自身および他人の知識が共有される．また,一連の操作は単一ブラウザ上で行えるように設計しているため，ユーザはシームレスに情報の取得ができるようになっている．

### 4.1 個人知識の構築

個人知識のモデル化は Mika が提唱した ACI モデルを参考にする[Mika 2005]．ACI モデルは，ソーシャルブックマークサービスのデータから，Actor（コンテンツにタグを付与したユーザ）-Concept（コンテンツに付与されたタグ）-Instance(タグが付与されたコンテンツ)から構成される（図 4）．

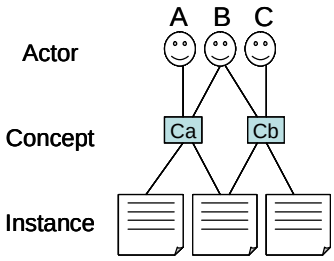


**図 4　ACI モデル**

この ACI モデルを本システムに適用すると，Actor（Web ページにマーキングを付与したユーザ）-Concept（Web ページに付与されたマーキング文字列（図 5 のグレー部分））-Instance（マーキングが付与された Web ページ）となる（図 5）．Concept は茶筌[松本 2003]でパースし，名詞および未知語と判定された語を用いる（図 5 の Ca1, Ca2）．図 5 の Concept-Instance 間にある点線は，システムのユーザ知識の獲得フェーズにおいて，ユーザが選択した Concept に対して表示された Web ページリンク一覧の中からユーザがリンク先として Instance を選んだことを示している．個人知識モデルとは図 5 における Actor の A･B･C それぞれの Concept および Instance へのリンク（実線および点線）構造のことを示す．故に図 5 は A･B･C の 3 ユーザの個人知識モデルが合わさったもので，この関係を用いて情報の表示フェーズで各ユーザに適した関連情報を推薦する．

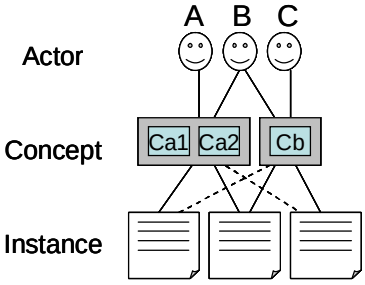


**図 5　個人知識モデル群**

## 5. まとめ

本論文では，個人独自の語の想起，すなわち個人知識に従って Web ページを分類することを目標とし，第 19 回人工知能学会全国大会で運用したイロノミーの分析結果から，Web ページへのマーキングの有効性とタグとしての問題点を得た上で，システムの設計を行った．システムは，ユーザがマウスカーソルで Web ページ内の文字列を選択すると同時に関連情報を表示することで，語の想起を促して個人知識を構築している．ユーザが本システムを使用し続けると，個人知識が構築されていき，ユーザの思考に沿った Web ページの分類が実現する．さらに，

ユーザの知識は Web ページ上にマーキングとして表示されて，ユーザ間で知識が共有される．これにより，他人の知識を通じて新しい情報を獲得できるようになった．

**謝辞**

本研究は情報通信研究機構の委託研究により実施したものである．

## 参考文献

[Mathes 2004] Adam Mathes : Folksonomies – Cooperative Classification and Communication Through Shared Metadata, LIS590CMC, 2004

[坂本 2006] 坂本 竜基, 中田 豊久, 伊藤 禎宣, 松岡 有希, 小暮 潔, 武田 英明：イロノミー：色付き傍線による Web 文章を対象としたフォークソノミー, JSAI2006, 2006

[土方 2002] 土方 嘉徳, 青木 義則, 古井 陽之助, 中島 周：マウス挙動に基づくテキスト部分抽出方式と抽出キーワードの有効性に関する検証, 情報処理学会論文誌, Vol.43, No.2, pp.566-576, 2002.

[鷹城 2002] 鷹城 徹, 武田英明：WWW ブラウジングを通した個人的知識の獲得と組織化, 電子情報通信学会論文誌, VOL.J85-D Ⅰ No.6 June 2002

[松岡 2006] 松岡 有希, 坂本 竜基, 中田 豊久, 伊藤 禎宣, 武田 英明: 論文概要に対する色付きアンダーライン付きステムの運用･分析, DEWS2006 , 2006.

[Salton 1991] Salton, G.: Developments in automatic text retrieval, Science, Vol. 253, pp. 974-980 (1991).

[野中 1996] 野中 郁次郎(著), 竹内 弘高 (著) , 梅本 勝博(翻訳) : 知識創造企業, 東洋経済新報社 ; ISBN: 4492520813 ; (1996/03)

[Mika 2005] Peter Mika: Ontologies are us: A unified model of social networks and semantics. Proceedings of the 4th International Semantic Web Conference (ISWC 2005), LNCS 3729, Springer-Verlag, 2005.

[松本 2003] 松本裕治他:「形態素解析システム『茶筌』version 2.3.3 使用説明書」, 奈良先端科学技術大学院大学, 2003.8