3D1-4

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

Ironomy: a Web pages forksonomy system utilizing colored underlines

坂本 竜基[*1] 中田 豊久[*2] 伊藤 禎宣[*1*3] 松岡 有希[*1*4*5] 小暮 潔[*1]
Ryuuki Sakamoto Toyohisa Nakada Sadanori Ito Yuki Matsuoka Kiyoshi Kogure

武田 英明[*1*4*5*6]
Hideaki Takeda

[*1]ATR メディア情報科学研究所
ATR Media Information Science Laboratories

[*2]北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology

[*3]東京農工大学
Tokyo University of Agriculture and Technology

[*4]総合研究大学院大学
Graduate University for Advanced Studies

[*5]国立情報学研究所
National Institute of Informatics

[*6]東京大学
Univ. of Tokyo

We propose a Folksonomy system, called Ironomy, which is aimed to classify Web pages. Folksonomy service is a new approach to classify large amount of information by adding metadata manually. This metadata called "Tag" are typed in by users as character strings. The proposed system uses, as tags, character strings in target Web pages which are underlined as results of a reading method with three colored underline. This system has been evaluated in JSAI2005.

1. はじめに

Web やメール等の日々追加・更新される膨大なデータからの必要情報を取得する情報検索技術は日常的生活になくてはならない存在となりつつある．ユーザの検索クエリーに従って言語処理等によって自動的に情報の濾過をおこなう検索エンジンはその代表例であろう．検索エンジンは，検索対象リソースが言語情報かつユーザが要求情報の定義を言語化可能な組み合わせが最も単純で効果的である．一方で，画像，音声といった非言語的リソースの検索は困難であるが，画像処理等によって対象リソースの検索キーとなる特徴を抽出して言語化することで言語情報の検索エンジンと同じフレームワークを適用するといった解決策が考えられる．このような計算機による自動処理を通して対象リソースを解析するアプローチとは逆に，対象リソースに機械的に生成困難な文字による二次情報を付与することで検索可能とするアプローチも考えられる．このような観点から，近年，画像やブックマークに対して人間の自由記述によるメタデータを付与して既存の検索エンジンでは困難な対象を直感的な語彙で検索可能とするサービスが提供されている．このようなサービスは，多くの人手で膨大な情報の分類をおこなうという意味から Folksonomy と総称される [Mathes, Golder 05] また，付与されるメタデータはタグと呼ばれる．

現在提案されている Folksonomy のサービスのうち，Web ページの URL をタグ付きで共有するサービスが有名である．これは，Web ブラウザにおけるブックマークを人間の言葉で表されたタグと共にサービスに投稿して，後に利用者はこのタグを検索したり，共通のタグが付与されている URL を辿ることによって既存の検索エンジンでは困難であった情報収集を目指したものである．画像等と異なりブックマークつまり Web サイトは言語情報であり検索エンジンの代表的な対象リソースであるため，さらにタグによる二次情報が有用であるという事実は直感的ではない．この理由としては，Folksonomy による情報検索は検索エンジンを用いたどちらかと言えば厳格な検索とは利用モチベーションが異なり，抽象的な語によるネットサーフィンやザッピングを目的とした場合が多いからではないかと考えられる．例えば，ブックマークを対象にした代表的サービスである del.icio.us(http://del.icio.us/) における 2006 年 4 月時点の popular tags 上位五個は「software」,「blog」,「music」,「news」,「design」と非常に抽象的である．これらのキーワードは対象リソースにも含まれている可能性が高いものの，既存の検索エンジンでは文章の一部にでも検索キーワードが含まれていれば検索結果とされるため有効な検索キーワードとは言えない．よって，抽象的なキーワードによる検索の場合，Folksonomy の自分や他人が Web 文章の特徴として明示するタグを検索するほうが検索精度が高いと考えられる．

一方で，一般的にメタデータは，その付与に関するコストの大きさが問題視されてきた．特に Folksonomy ではタグを手入力する必要があるため利用者に大変な手間を強いることになる．代表的存在である del.icio.us や Flikr(http://www.flickr.com/) の利用人数の多さやタグの多さを鑑みると手入力によるコストの問題はそれほど深刻ではないとも考えられるが，多くのサービスのタグ入力フォームには補完機能が実装されていることからも，手間を減らす機能はユーザビリティ向上の面で有益であると考えられる[*1]．

本稿では，自由入力によるタグではなく，文章に含まれる文字列をタグとして考える Web 文章の分類を目的とした Folksonomy システム（以下，イロノミーと呼ぶ）を提案する．Flickr 等が対象とする画像とは異なり，本システムや del.icio.us 等が対象とする Web 文章は文字列を包含している．よって，もし分類するに足る特徴的なキーワードをその文章から抽出することができれば，そのキーワードはタグとして機能すると考えられる．イロノミーは，Web 文章の一部に対するマウス選択

連絡先: 坂本 竜基, 国際電気通信基礎技術研究所, 住所：〒619-0288 京都府相楽郡精華町光台 2-2-2 ATR メディア情報科学研究所, 電話：0774-95-2553, Fax：0774-95-1408

[*1] この補完機能は，手間の低減と同時に Folksonomy で指摘されている表記ゆれ対策としても機能するとも考えられる．

がユーザの知的興味を反映した重要文の抽出に有用であるという知見 [土方 02, 鷹城 02] に基づき，マウス操作によって選択されて引かれた傍線（下線）が指示する文字列をその文章の特徴的なキーワード，つまりタグとみなす．この下線は，三色ボールペン読書法 [齋藤 02] と呼ばれる傍線を引きながら文章を読み進める手法の結果として保存されるもので，ユーザは，単に三色ボールペン読書法による文章読解を進めるだけで，タグ付けを意識することなく Folksonomy サービスを利用可能となる．

以下，三色ボールペン読書法の概要とイロノミーのシステム説明をおこなった後，2005 年度人工知能学会の Web サイトにおける発表情報を載せた Web ページを分類するシステムとしてイロノミーを運用した結果を紹介する．

2. 三色ボールペン読書法

齋藤は，重要な箇所に 3 種類の色付きの傍線を付与しながら文章を読み進めていく読書法を提案しており，これにより文章の理解が深まるとされている [齋藤 02]．この読書法では，下線の色にそれぞれ意味が与えられており，赤色は「客観的にとても重要」，青色は「客観的にまあ重要」，緑色は「主観的に重要」と区別している．傍線は，同じ箇所に別の色を同時に付加してもよいとされているが，イロノミーでは HTML の制約から一つの文字列に対して色は排他である．

3. イロノミー

イロノミーは，三色ボールペン読書法による傍線入力を支援する入力インタフェース部と，付帯された傍線から Web ページの整理，分類の支援する検索推薦部に別れる．以下，各モジュールについて説明する．

3.1 入力インタフェース部：HTML 文章への三色ボールペン読書法支援

マウスで HTML の一部分を選択することによって三色ボールペン読書法を支援するユーザインタフェースを提供した．入力インタフェース部は，まず，傍線としてマウスによって選択された文字列の HTML 上での位置特定をおこない，これを日時，ユーザ ID，ユーザによって選択された線の色と共にサーバに送信する．選択文字列の位置特定は DOM2 の Range によりおこなわれ，具体的な位置情報としては選択文字列の親ノードまでの XPath，選択文字列，文字列の登場順位が保存される．傍線の情報は該当 HTML が読み込まれたタイミングでサーバから送信され，入力時とは逆の順番で傍線の位置を特定し，線の色を指定した上で CSS（Cascading Style Sheets）によって傍線の描画をおこなう．このインタフェースのスクリーンショットを図 1 に示す．

3.2 検索推薦部：傍線による Web ページの整理・分類

三色ボールペン読書法用インタフェースによって付与された各下線は，専用フォームにおいて検索される．ユーザは，一つ以上検索したい線の色及び自分の線か他人の線かをチェックボックスで指定した上で文字列入力フォームに検索したいタグ（傍線が引かれた文字列）を入力して検索する．図 2 の上段にこの検索用フォームを示す．

イロノミーには，このようなユーザが能動的に情報検索をおこなうだけでなく，ユーザ自身の下線から他の未読文章が推薦される，受動的な情報取得を支援する機能も存在する．図 2 の下段は推薦結果を表示した例であり，この結果は中段のボタンを押下することで得られる．

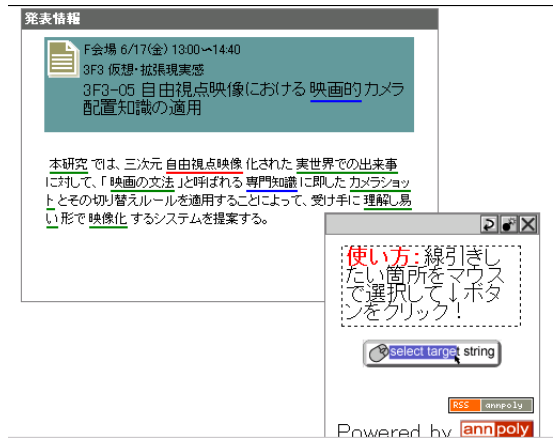

図 1: HTML への 3 色ボールペン読書法支援インタフェース

推薦アルゴリズムは，現在のところ単純に GroupLens 等と同じくピアソンの積率相関係数によるユーザ間の類似度を用いて対象ユーザが未読の各文章のうち評価が高いものを推薦する方式を採用している．具体的には，各文章に引かれた線の数を評価とみなし，ユーザ u_i とユーザー u_j 間の類似度を以下の式で求める．

$$sim(u_i, u_j) = \frac{\sum_{c \in A_i \cap A_j} (v_{i,c} - \bar{v_i})(v_{j,c} - \bar{v_j})}{\sqrt{\sum_{c \in A_i \cap A_j} (v_{i,c} - \bar{v_i})^2 (v_{j,c} - \bar{v_j})^2}} \quad (1)$$

ここで，ユーザ u_i に対して A_i は評価した文章の集合，$v_{i,c}$ は文章 c への評価値，$\bar{v_i}$ は評価の平均とする．

次に，u_i が未読の文章 A_n に対する予想評価 $s_{i,n}$ を以下のように計算する．

$$s_{i,n} = \bar{s_i} + \frac{\sum_{u_n \in U_n} sim(u_i, u_k)(v_{k,n} - \bar{v_n})}{\sum_{u_n \in U_n} |sim(u_i, u_k)|} \quad (2)$$

ここで，U_n は A_n を評価したユーザの集合とする．以上の計算は線の色別におこなわれ，各色における重要度が上位のものを数件推薦する．評価の入力を色毎に分けることは，三色ボールペン読書法の手法から考えて，客観的な重要度でみた場合の推薦と主観的なそれに分けて推薦することに繋がり，より細かな情報分類の実現が期待できる．

4. 運用実験

本章では，第 19 回人工知能学会全国大会[*2]において運用された学会サポートサービス [武田 06] の 1 サービスとしてイロノミーを運用した結果を述べる．サポートサービスでは，ユーザのスケジュール管理や全 297 件の発表のタイトル，著者，概要などを閲覧できた．発表は 1 件毎に専用ページが用意されていたため，これをイロノミーの分類対象とすることにした．

3 日間の大会会期中にイロノミーを利用したユーザ数は 27 名であり，下線は 265 本引かれた．下線の色別の内訳は，赤色の線は 83 本，青色が 101 本，緑色が 81 本であり，すべての色の下線が満遍なく引かれていたことが判る．

以上の入力インタフェース部の利用により付与された下線情報を検索推薦部において活用した結果としては，まず，検索に

*2 http://www-kasm.nii.ac.jp/jsai2005/schedule/

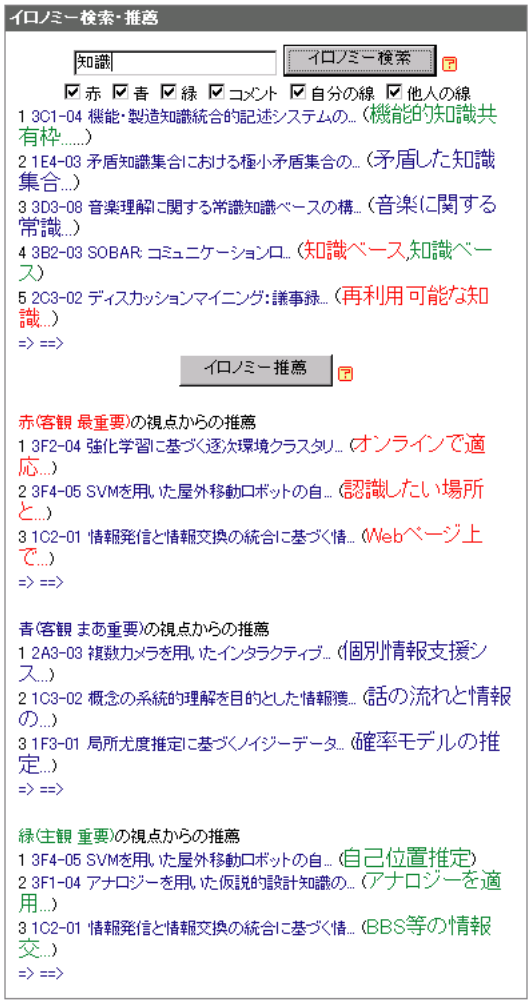

図 2: 上部：傍線（タグ）の検索用フォーム 下部：色別の推薦結果

より表示された発表ページ一覧のうちの一つをクリックして該当発表ページへジャンプした（リンクアンカをクリックした）件数は 260 回であった．また，協調フィルタリングによる推薦結果をつかってその発表ページへジャンプした件数は 160 件認められ，内訳は赤色が 82 件，青色が 42 件，緑色が 36 件であった．赤色が多いことから推薦においてユーザは客観度を重視した傾向が伺える．

学会終了後におこなった，学会サポートサービス全体で利用者に対する Web フォームによるアンケート調査の結果を報告する．全 107 名のアンケート回答者のうち 21 名がイロノミーを利用したと回答した．まず，サービスの印象を 5 段階（「そう思う」「すこしそう思う」「どちらともいえない」「あまりそうは思えない」「そう思わない」）で評価する質問を 3 種類用意した．

「このサービスは便利だった」という質問に対しては，「そう思う」から順に 4 名，5 名，7 名，4 名，1 名という回答結果であった．「このサービスは面白かった」という質問に対して，6 名，6 名，5 名，3 名，1 名であり，「このサービスを来年も利用したい」に対しては 7 名，4 名，6 名，3 名，1 名という回答結果であった．これらの回答結果は，利便性について意見が分かれているものの，システムの可能性についてはポジティブであったことを示している．

次に，三色ボールペン読書法に対する理解を調査する目的で「色の下線を引く際，それぞれの色の意味を理解していましたか？」という質問をおこなった．これに対して，「はい，理解していました」という項目に 12 名，「いいえ，理解していませんでした」に 3 名，「理解していましたが，色分けするつもりはありませんでした」に 6 名が投票した．この結果は，半数以上のユーザが色付き傍線の意味を理解して利用していたものの，十分に意図が伝わらないユーザも存在していたことを示唆している．

また，イロノミーのどの部分に有用性を感じるのかを調査する目的で，「Web ページに下線を追加できることについて該当するものにチェックしてください.」という質問をおこなった．回答項目は，我々が考えた機能的特長を列挙したリストに複数回答可でチェックをつける形式とした．これに対して，「メモを Web ページに残せて便利である」に 9 名，「他人の書いた下線が役に立つことがある」に 11 名，「気になるページに下線を引き，後で検索できることが便利である」に 5 名がチェックした．この結果は，自分で引いた下線の利用もさることながら，他人の下線にも興味があったことを示している．

最後に自由記述でシステムに対するコメントを記述する欄では，「３色の使い分けが意外としにくかった． 線引きを自分の習慣にして，大量の下線（＋アノテーション）データを引くことができれば非常に強力なツールとなると思うが，問題はどのようにして習慣にするか.」「削除の仕方がわからない」「概要ではなく，本文に下線をひきたかった」等の意見がみられた．これらは操作やシステムの仕様に関する事柄であり，操作のインストラクションや機能拡張については今後の課題とする．

5. まとめ

本稿では，Web ページに含まれる文字列情報をタグとする Folksonomy を提案し，三色ボールペン読書法の色付き傍線が指示する文字列を入力とするシステムについて述べた．また，JSAI2005 において運用実験をおこなった結果を紹介した．今後は，現在は下線の本数によって評価されている協調フィルタリングのスコアリングを本来ユーザが対象 Web ページに対して評価したいスコアとして近似する手法に代え，推薦精度の向上に努めたい．

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

[Golder 05] Golder, S. and Huberman, B. A.: The Structure of Collaborative Tagging Systems, Technical report, Information Dynamics Lab, HP Labs (2005)

[Mathes] Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata, *http://www.adammathes.com/academic/computer-mediated- communication/folksonomies.html*

[齋藤 02] 齋藤 孝：三色ボールペンで読む日本語，角川書店 (2002)

[鷹城 02] 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 D-I, Vol. J85-D-I, No. 6, pp. 549–559 (2002)

[土方 02] 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)

[武田 06] 武田 英明, 西村 拓一, 松尾 豊, 濱崎 雅弘, 藤村 憲之, 石田 啓介, ホープ トム, 中村 嘉志, 沼 晃介, 永田 寛, 中川 修, 新堀 英二, 藤吉 賢, 坂本 和彌, 高橋 徹, 坂本 竜基：JSAI2005/UbiComp05 におけるイベント空間情報支援システムの開発・運用, 人工知能学会第 20 回全国大会 (2006)