マーキングを用いたソーシャルタギングの有効性に関する検証

松岡有希[†1,†2] 坂本竜基[†3] 伊藤禎宣[†3,†4]
大向一輝[†1,†2] 武田英明[†1,†2,†3] 小暮潔[†3]

近年，ソーシャルブックマークのタグを Web ページへのメタデータとして利用することが注目されている．メタデータは誰が見ても Web ページの内容を把握できるように，Web ページの内容が直接反映された語で書かれることが望ましい．そこで本研究では，ユーザが Web ページの内容と直接関係のあるタグを付与できるようにするため，Web ページ内の文字列に下線やハイライトマーカを付与するマーキングに着目した．我々はマーキングによるソーシャルタギングの有効性を検証するために，人工知能学会全国大会で運用したマーキング共有システムで得られたデータを使って分析をした．Web ページ上にマーキングを付与できるシステム運用で得られたデータの分析によると，tfidf 値の高い語に下線が付与される可能性が高いということが分かった．次に，マーキングを利用してページ推薦を行うシステムを運用したところ，ユーザは tfidf を用いたページ間類似度によるページ推薦よりも，他のページに付与されているマーキング文字列内の語を使ったページ推薦を選択することが示唆された．ユーザは統計的手法で得られる文書内の特徴語よりも，人がマーキングを付与した語を特徴語とした Web ページを選択することが示唆されたので，マーキング文字列内の語をタグと見なす手法が機能する可能性が見出された．

Effectiveness of Social Tagging Based on Marking

Yuki Matsuoka,[†1,†2] Ryuuki Sakamoto,[†3] Sadanori Ito,[†3,†4]
Ikki Ohmukai,[†1,†2] Hideaki Takeda[†1,†2,†3] and Kiyoshi Kogure[†3]

Recently, tags of social bookmark services are used as metadata of web pages. In this case, meta data are desirable to be written in words which are related with contents of web pages. In our research, we focus marking which users underline or highlight characters in web pages to make users can add tags which are related with contents of web pages. We investigated which words users underlined in web pages and whether users selected web pages added tags by marking with operation systems at a conference. According to the analysis of user logs based on a marking system, we found that users have the potential to underline words with high tfidf value. And, according to the analysis of user logs based on a recommendation system, users selected web pages recommended by using words in marked characters to web pages recommended by using page similarities based on tfidf value. Therefore, we found that users select words in marked characters to words calculated by a statistical method. As results, we believe it is effective to extract some words in marked characters to be tags.

1. はじめに

セマンティック Web は，Web ページの内容を人間だけでなく，機械が理解できるようにすることを目標とした技術である[1)]．機械が Web ページを理解できるようにするためには，Web ページの内容を記述したメタデータが必要となる．従来，考えられてきた Web ページに対するメタデータの記述方法[10)~12)] は，最初にオントロジーを RDFS [2)] や OWL [3)] で準備してから，そのオントロジーを使って Web ページにメタデータを記述するというトップダウンな方法である．しかし，WWW では様々なドメインを扱っているため，各ドメインに対応したオントロジーを用意するのは簡単ではない．また，人は各ドメインに対してそれぞれ異なる見方をするため，一意のオントロジーを定義するのは困難である．もしオントロジーが用意できても，Web ページで扱われる言葉の変化についていけない場合がある．また，オントロジーを使ってメタ

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

データを記述するのは専門的な知識が必要とされる．これらの問題より，これまでのメタデータの記述方法ではセマンティック Web を実現するのは難しい．

一方で，ソーシャルブックマークのタグをメタデータとして利用することが注目されている．ソーシャルブックマークは，ユーザが “タグ” と呼ばれるキーワードとともに Web ページをブックマークし，複数のユーザ間でブックマーク情報を共有するサービスである．タグは，ユーザが Web ページを整理したり，思い出しやすくしたりするために Web ページに与えるキーワードによる説明である．ユーザは，自由な言葉を使ってタグを付与したり，1 つの Web ページに対して複数のタグを付与したりすることができる．この方法によってメタデータを用意する利点としては以下の点があげられる．

- ソーシャルブックマークでタグを付与することは簡単なので，オントロジーの専門知識を持っていない一般ユーザが大勢参加できる．
- タグからオントロジーを抽出することで，Web ページで使われる言葉の変化に対応することができる．

しかし，ソーシャルブックマークのタグはユーザが自由に言葉を与えることができるため，ページの内容とは直接関係のないタグが生成されることがある．そういったタグは，セマンティック Web のメタデータとして Web ページの内容を記述する際にノイズとなる．そこで本研究では，ユーザが Web ページ内の文字列に下線やハイライトマーカを付与するマーキングに着目した．本稿では，実際に運用したマーキング共有システムで得られたデータを使って，マーキングによるタグ付けがソーシャルタギングとして機能するかどうかを調べた．

以下，2 章では関連研究について，3 章ではマーキングによるタグ付けに関する趣旨について，4 章ではマーキングが付与された語の特徴について，5 章ではマーキングによるタグ付けの有効性について，6 章ではまとめと今後の課題について述べる．

2. 関 連 研 究

Mika[4] や Wu[5] らは，ソーシャルブックマークのタグとタグを付与したユーザ，タグが付与されている Web ページの 3 つの関係を使って，タグ間の関連性を発見している．Mika は，タグ間の上位・下位概念を見つけることによって，ライトウェイトオントロジーを作った．Wu は，概念が似ている語集合を見つけ出した．Specia ら[8] や Damme ら[9] は，フォークソノミーとセマンティック Web の統合に向けて，タグに意味を付加するというアプローチをとっている．Specia は，既存のオントロジーのコンセプトやプロパティやインスタンスにタグをマッピングしたり，マッピングされたタグ間の関係を決定したりする．Damme は，ソーシャルブックマークのタグだけでなく，WordNet[☆] や Wikipedia[☆☆] といった辞書や RDF や OWL で書かれたオントロジーを利用して，タグとそれらをマッピングすることでオントロジーの生成を試みている．しかし，彼らは Web ページに付与されているタグがページの内容を表す語であるかどうかの考慮はしていない．

一方で，Web ページに自動的にメタデータを付与する研究もある．Artequakt システム[7] は，Web ページから自動的に知識を抽出して，オントロジーを表示する．Cimiano ら[6] は，Web ページ内の名詞をクエリとし，Google API の検索結果を使って，コンセプトとインスタンスを見つけ出している．しかし，我々はユーザが Web ページ内の語を選択すること（マーキング）によって得られた語，すなわち Web ページ内の必要とされる語のみをメタデータ化することを考えている．

3. マーキングによるタグ付け

ソーシャルブックマークでは，ユーザは Web ページにタグを自由に付与することによってソーシャルタギングを行う．タグは，下記のように分類することができる[13]．

(1) Web ページの主題に関すること
(2) Web ページに書かれている内容の種類
例：article，blog，book
(3) Web ページを作成した人の名前
(4) 単独では意味がなく，分類のためのタグ
例：丸めた数字，記号
(5) タグを付与したユーザの意見を反映した形容詞
例：scary，funny，stupid
(6) Web ページとタグを付与したユーザの関係
例：mystuff，mycomments
(7) Web ページに対するユーザのタスク
例：toread，jobsearch

(4)〜(7) のような個人的な意見や解釈が反映されたタグがブックマークされている Web ページを見ても，タグを付与したユーザ以外は期待どおりの情報を獲得しにくい[14]．たとえば，funny や toread といっ

☆ http://wordnet.princeton.edu/
☆☆ http://wikipedia.org/

たタグは Web ページに対する評価や重要度がユーザによって異なるので，これらのタグにブックマークされている Web ページを見ても役に立たないユーザがいる．このように，ソーシャルブックマークにおいてユーザが自由な言葉で付与したタグの中には，タグの内容に即した情報を取得したい場合に適していないものもある．したがって，(4)〜(7) のようなタグはページの内容を直接反映させたタグとはいえない．一方で，(1)〜(3) のようなタグは Web ページの内容と直接関係するタグのため，ユーザがこれらのタグにブックマークされている Web ページを見たとき，タグの内容に即した Web ページを取得できる．

我々は Web ページの内容を直接反映したタグを生成するために，ユーザが文章内の文字列に下線やハイライトマーカを付与するマーキングに着目した．マーキングは書籍を読むときに文章内の文字列に下線を引いたりハイライトさせたりする行為であり，多くの人にとって馴染み深い行為である．すなわち，ユーザが Web ページ内にマーキングを付与した語や文字列をタグと見なす．このとき，マーキングによるタグ付けの有効性を検証するために，マーキングされた語の性質の分析やマーキングによるタグが付与されている Web ページとそうでない Web ページのどちらが選択されるかの分析を行う．なお，本研究では，マーキングングによるタグ付けがソーシャルタギングとして機能するかどうかの調査を目的としており，マーキングによるタグ付けが既存のソーシャルブックマークと比較して優れているということを示すわけではない．以降，4 章においてマーキングされた語の特徴について，5 章においてマーキングによるタグ付けのソーシャルタギングとしての有効性ついて検証した．

4. マーキングが付与された語に関する分析

本章では，マーキングが付与された語に関する分析について述べる．我々は分析のために，2005 年 6 月 15 日（水）から 6 月 17 日（金）に開かれた第 19 回人工知能学会全国大会（JSAI2005）で運用された “イロノミー” で得られたデータを利用した．

4.1 システム概要

イロノミーは，発表ページ（学会で発表される論文の情報が書かれたページ）内の論文概要の文章に対し，三色ボールペン読書法[16] に従ってユーザが色付きの下線を付与できるシステムである[15]．三色ボールペン読書法は，客観的にとても重要だと思う箇所を赤色で，客観的にまあ重要だと思う箇所を青色で，主観的に重要だと思う箇所を緑色で下線を引きながら読書をする方法である．イロノミーは，ユーザが三色ボールペン読書法と同じように，論文概要の文章内の客観的または主観的に重要な箇所に赤・青・緑の 3 色を使って下線を付与することができる機能を提供している．ユーザは**図 1** 内の「概要にイロノミー用の下線を引く」というリンクボタンをクリックすることで，概要内の好きな箇所に三色ボールペン読書法を用いて色付きの下線を付与することができる．

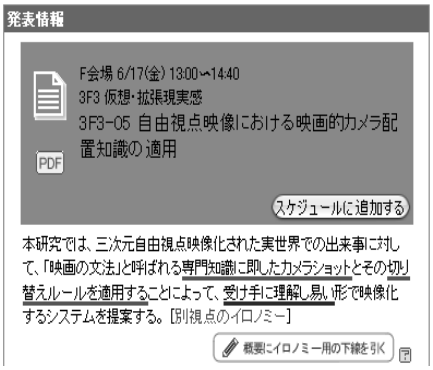


図 1 イロノミーの利用よって下線が付与された発表ページの画面
Fig. 1 Screen of a presentation page added underlines by using ironomy.

4.2 運用結果

イロノミーが対象とした Web ページは，学会で発表される論文の概要文が書かれた発表ページで，全部で 294 ページあった．運用の結果，イロノミーを使用したのは開発者を除いて 27 人だった．ユーザが付与した下線の数の平均本数は 6.2 本で，分散は 87.2，標準偏差は 9.3 だった☆．下線が付与された論文概要は 67 個あり，下線の総数は 168 本で，赤線の数が 47 本，青線の数が 64 本，緑線の数が 57 本だった．

4.2.1 下線が付与された語に関する分析

本章では，ユーザが Web ページ内で下線を付与した語にどのような特徴があるのかについて，イロノミーの運用で得られたデータを使って分析を行う．我々は，ユーザがどのような語に下線を付与したのかについて調べるために，文書内の語を統計的に特徴付ける手法として広く用いられている tfidf[20] を使う．文書内の語は，tfidf で求めた値を使うことにって，下記のように特徴付けることができる．

- tfidf 値が高い語は，対象文書内で出現頻度が高く，他の文書には現れにくいので，対象文書の特

☆ 今回の実験は，被験者の属性の制御や統制が不十分な環境で行っており，分散が大きい．また，全被験者がどの程度三色ボールペン読書法を理解してマーキングをしたかも不明であり普遍性があるデータとはいえないが，1 つの傾向として報告する．

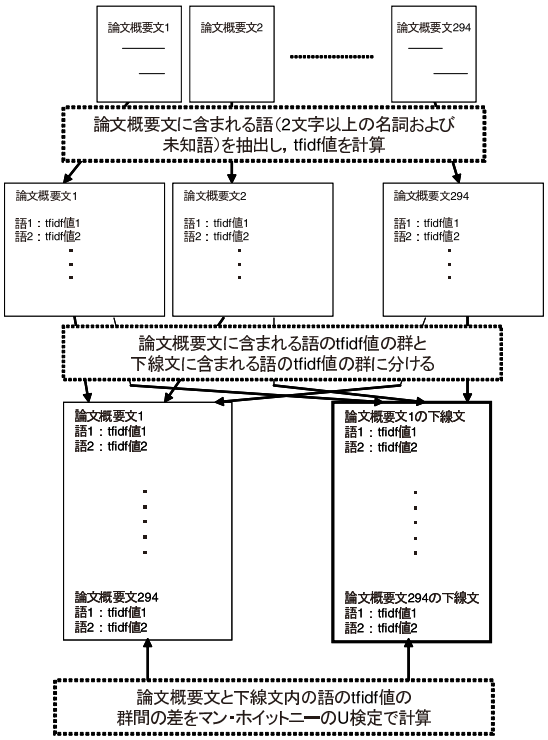


図 2 分析手法
Fig. 2 Method of analysis.

徴語といえる．

- tfidf 値が低い語は，対象文書内での出現頻度が低く，他の文書に頻繁に出てくる語であるため，一般語である可能性が高い．

本項では，ユーザが下線を付与した語の tfidf 値には，どのような傾向があるのかについて調べるために，論文概要文に含まれる語と下線文に含まれる語を tfidf 値を用いて比較した．分析手法を**図 2** にまとめた．論文概要文に含まれる語は，各論文概要文ごとに茶筅[17]を使って形態素解析をし，2 文字以上の名詞および未知語を採用した．tfidf 値は，この語を使って式 (1) により求めた．

$$tfidf(w,a) = tf(w,a) \cdot idf(w) \qquad (1)$$

$tf(w,a)$ は論文概要文 a における語 w の出現回数を，$idf(w)$ は語 w が全論文概要文のうちどのくらいの頻度で出現するかの尺度であり，$\log(N/df(w))$ で求める．$df(w)$ は語 w が含まれる論文概要文の数を表す．N は論文概要文の総個数を表し，今回は 294 個である．下線文に含まれる語は，下線文に元の論文概要文内の語がある場合，その語を下線文に含まれる語として採用した．また，下線文に含まれる語の tfidf 値は，元の論文概要文に含まれる語の tfidf 値をそのまま利用した．**表 1** はすべての論文概要文と全ユーザによる

表 1 全論文概要文と全下線文に含まれる語に関する tfidf 値
Table 1 Tfidf value of word in all abstracts and underlines.

	全論文概要文	下線文
平均	4.5	5.9
分散	8.9	16.8
語数	6481	456

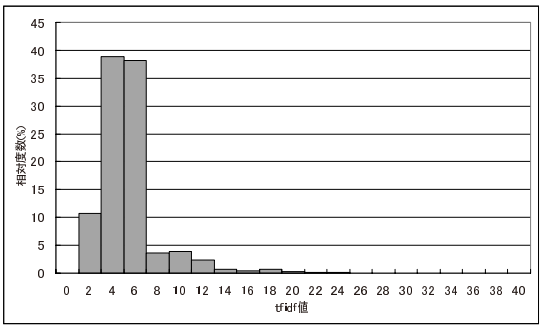


図 3 全論文概要文に含まれる語の tfidf 値のヒストグラム
Fig. 3 Histogram of tfidf value in all abstracts.

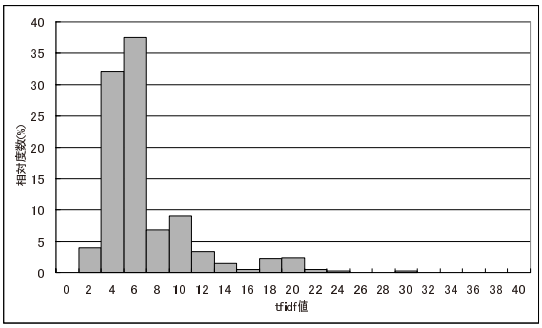


図 4 全下線文に含まれる語の tfidf 値のヒストグラム
Fig. 4 Histogram of tfidf value in all underlins.

下線文に含まれる語の tfidf 値群の平均値と分散，語数を示している．全論文概要文に含まれる語の tfidf 値群と全ユーザによって付与された下線文に含まれる語の tfidf 値群とで平均値の差がないかを調べるために，マン・ホイットニーの U 検定を行った．

検定の結果，$P < 0.05$ となり，両者に有意差があるという結果が得られた．**図 3** は全論文概要文に含まれる語の tfidf 値のヒストグラムで，**図 4** は全下線文に含まれる語の tfidf 値のヒストグラムである．図 3, 4 において，tfidf 値のデータ区間が 6 以降の語の相対度数を見比べると，論文概要文に含まれる語の相対度数よりも下線文に含まれる語の相対度数のほうが高いことが分かる．したがって，ユーザが付与した下線文は tfidf 値の高い語，すなわち Web ページ内の特徴語を多く含む傾向があることが分かった．

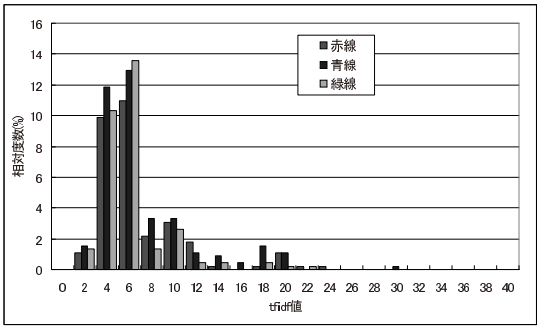


図 5 各色の下線文に含まれる語に対する tfidf 値のヒストグラム
Fig. 5 Histogram of tfidf value in each color underline.

表 2 各色の下線文に含まれる語に関する tfidf 値
Table 2 Tfidf value of words in all underlines.

	赤線	青線	緑線
平均値	6.0	6.3	5.2
分散	17.9	20.7	10.5
語数	141	174	141

4.2.2 色の付け方に関する分析

本項では，全ユーザによって付与された下線文のうち，それぞれの色線に含まれる語の tfidf 値がどのような傾向を示すのかについて調べた．**図 5** は，赤・青・緑色の下線文に含まれる語の tfidf 値のヒストグラムである．**表 2** は，各色の下線文に含まれる語の tfidf 値の平均値と分散，および語数を示している．赤・青・緑色の下線文に含まれる語の tfidf 値群の平均値に差があるかどうかを，クラスカル・ウォリス検定で調べた．その結果，$P > 0.05$ となり，群間の有意差を得ることはできなかった．一方で，各ユーザごとの色の付け方を調べると[18]，tfidf 値の高い語に赤色および青色の下線を付与するユーザもいれば，緑色の下線を付与するユーザもおり，色の付け方はユーザによって様々である．これは人によって主観および客観の判断が異なるからと思われる．したがって，各ユーザによって色の付け方は異なるが，ユーザ全体でみると平均化されて各色線に含まれる語の tfidf 値の有意差がなくなると考えられる．4.2.1 項で得られた結果と合わせると，ユーザ全体でみると色にかかわらず tfidf 値の高い語に下線が付与される可能性が高いということが分かった．

4.2.3 下線が付与された語とソーシャルブックマークタグの分類

本項では，ユーザが下線を付与した語が 3 章で紹介したソーシャルブックマークタグの分類のうち，どの分類に該当するのかについて述べる．タグの分類のうち (4)〜(7) は，個人的な意見や解釈が反映されたタグに関する分類であった．下線を付与することというのは著者が書いた語を選択することなので，下線を付与した語に個人的な意見を反映することはできない．したがって，下線が付与された語は (4)〜(7) の分類にはあてはまらない．

次に，下線が付与された語は，Web ページの内容と直接関係するタグである (1)〜(3) の分類に該当するのかについて述べる．今回は学会で発表された論文の概要文を対象としたので，下線が付与された語の中に (3) の Web ページを作成した人の名前にあたる語はなかった．しかし，著者名が書かれている文書，たとえば，論文そのものやプロフィールが書かれた Web ページを対象にした場合は，人名に下線が付与される可能性がある．(2) の Web ページに書かれている内容の種類に関する分類においても，対象文書に “論文” という語は書かれていなかったため，該当する語はなかった．この場合，論文や Web ページに文書の属性が書かれていれば，下線が付与される可能性がある．

最後に，下線が付与された語は (1) の Web ページの主題に関する語に該当するのかについて述べる．下線が付与された回数が多い上位 10 個の語は，順に，情報，ネットワーク，知識，状況，研究，行動，提示，クラスタリング，ユーザ，パターンだった．これらの語は論文概要文に含まれる語なので，論文の内容と関係のある語，とはいえるが，主題に関する語であるとはいいきることはできない．そこで，これらの語が論文の主題を表すタイトルとして使用されているかどうかを調べた．論文のタイトルに含まれる回数が多い上位 30 個の名詞のうち，情報，ネットワーク，知識，行動，クラスタリング，ユーザの 6 語が含まれていた．よって，(1) の分類に該当する語があることが分かった．これらの考察より，下線が付与された語は (1)〜(3) の分類に該当する可能性があることから，Web ページの内容と直接関係するタグとして利用できるものと思われる．

5. マーキングによるタグ付けの有効性に関する分析

本章では，ユーザがマーキングを付与した語や文字列をタグとした場合，マーキングによるタグ付けがソーシャルタギングとして機能するかどうかを調べた．我々は，2006 年 6 月 7 日（水）から 6 月 9 日（金）に開かれた第 20 回人工知能学会全国大会（JSAI2006）で運用された “合口” で得られたデータを用いて調べた．

5.1 システム概要

合口は，マーキングを利用して発表ページを推薦す

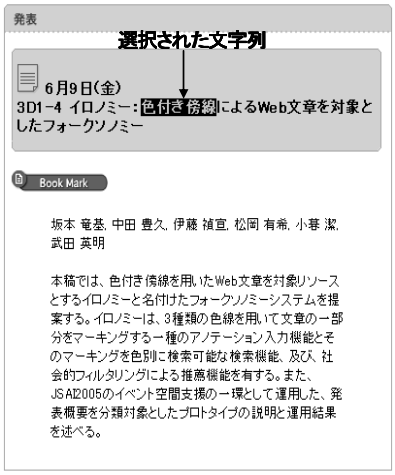


図 6 ユーザが Web ページ内の文字列を選択する
Fig. 6 Users select a string in a web page.

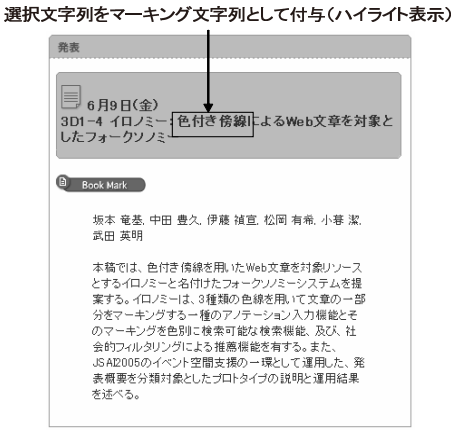


図 8 選択文字列をマーキング文字列として発表ページ上に付与する
Fig. 8 The system adds the marking string on the web page.

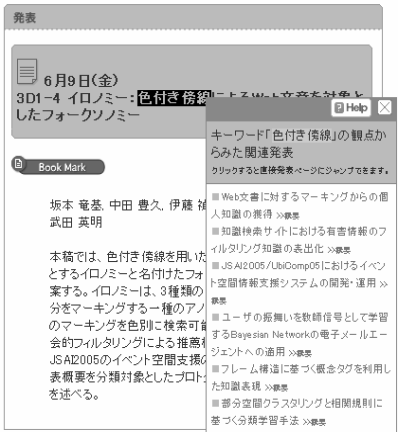


図 7 推薦リンクが書かれた小窓を表示する
Fig. 7 The system pops out the window displayed recommendation links.

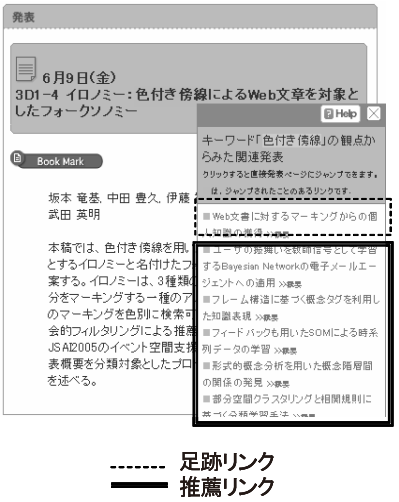


図 9 足跡リンクと推薦リンクが書かれた小窓を表示する
Fig. 9 The system pops out the window displayed footprint links and recommendation links.

るシステム[19] である．合口はユーザが発表ページ内の文字列をマウスカーソルで選択した際（**図 6**），推薦アルゴリズムに基づいて算出された他の発表ページへの推薦リンク（発表ページのタイトル）が書かれた小窓を表示する（**図 7**）．ユーザは提示された推薦リンクの中から気に入ったものをクリックをすると，クリック先の発表ページへ遷移すると同時に，合口はユーザが選択した文字列をマーキング文字列として発表ページ上に付与する（**図 8**）．発表ページに付与されたマーキング文字列は，ハイライト表示されて他の発表ページへのリンクアンカの役割を果たす．このマーキング文字列をユーザがマウスカーソルでなぞると，合口は足跡リンクと推薦リンクが書かれた小窓を表示する（**図 9**）．足跡リンクは，以前誰かがこのマーキング文字列（選択文字列）から遷移したことのある他の発表ページへのリンクで，推薦リンクは推薦アルゴリズムに基づいてシステムが推薦した他の発表ページへのリンクである．

ここでは，ユーザがマーキングを付与した文字列内の語をタグと見なした場合のソーシャルタギングとしての有効性を検証するために，4 種類の推薦アルゴリズムを使った他ページへの推薦機能を実装した（図 7）．また，ユーザがマーキングを付与した文字列をタグと見なした場合のソーシャルタギングとしての有効性を検証するために，足跡リンクと推薦リンクを同時に表示するよう実装した（図 9）．

5.2 推薦アルゴリズム

4 章より，tfidf 値の高い語はユーザがマーキングする語である可能性が高いということが分かった．そこで我々は，tfidf 値の高い語によるタグと，ユーザがマーキングを付与した語によるタグのどちらが付与されている Web ページを選択するのかについて調べるために，4 種類の推薦アルゴリズムを用意した．用意した推薦アルゴリズムは下記のとおりである．

A) tfidf を使ったページ間類似度による推薦

B) 発表ページにユーザが付与したマーキング文字列の数を使った協調フィルタリングによる推薦

C) ユーザがマウスカーソルで選択した文字列内の語と他の発表ページに付与されているマーキング文字列内の語とのマッチングによる推薦

D) ユーザがマウスカーソルで選択した文字列内の語と他の発表ページ内の語とのマッチングによる推薦

アルゴリズム A では，ユーザがマウスカーソルで文字列を選択した発表ページに対して，tfidf を使ったページ間類似度が高い他の発表ページを推薦する．ページ間の類似度は，発表ページ内の文章を茶筅[17] を使って形態素解析し，得られた語の中から 2 文字以上の名詞および未知語の tfidf 値を求めた後，発表ページ間の類似度をコサイン類似度を使って計算した．アルゴリズム A は，ユーザが選択した文字列や他のページに付与されているマーキング文字列とは関係なく，ユーザがマウスカーソルで文字列を選択した発表ページに対して，類似度の高い発表ページを推薦する．

アルゴリズム B では，協調フィルタリング[21] を用いて，似た嗜好を持ったユーザが選択したことのある発表ページを推薦する．ユーザがマーキング文字列を付与した発表ページのうち，同じ発表ページにマーキングを付与したことのあるユーザ同士は似た嗜好を持っている可能性が高い．そこで，ユーザによる発表ページへの評価値を，発表ページ上にユーザが付与したマーキング文字列の数とし，協調フィルタリングによる計算を行った．アルゴリズム B は，ユーザがマウスカーソルで選択した文字列は用いずに，他のページに付与されているマーキング文字列を用いるが，マーキング文字列内の語はいっさい考慮しない．

アルゴリズム C では，ユーザが発表ページ内で選択した文字列内の語と他の発表ページに付与されているマーキング文字列内の語とのマッチングを行ってマッチした発表ページを推薦する．他の発表ページ上に付与されているマーキング文字列は全ユーザによって付与されたものを対象とする．また，マッチングに利用する語は，選択文字列内およびマーキング文字列内の名詞および未知語である．

表 3 各推薦アルゴリズムおいて使用する文字列の比較
Table 3 Comparison of used characters in each recommendation algorithm.

推薦アルゴリズム	選択文字列	マーキング文字列	発表ページ内の文字列
A	×	×	○
B	×	○	×
C	○	○	×
D	○	×	○

アルゴリズム D では，ユーザが発表ページ内で選択した文字列内の語と他の発表ページ内の語とのマッチングを行ってマッチした発表ページを推薦する．アルゴリズム D は，ユーザが発表ページ内で選択した文字列内の語を検索クエリとし，他の発表ページ内に含まれているかどうかを調べている．一般的にユーザが Web ページを探すのに最も利用するのは検索エンジンであるため，検索エンジンで行われることと同じ手法を推薦に取り入れた．

各推薦アルゴリズムにおいて，選択文字列（ユーザがマウスカーソルで選択した文字列）や発表ページに付与されたマーキング文字列，発表ページ内の文字列を使用するかどうかを**表 3** にまとめた．

アルゴリズム A のページ間の類似度はシステムの運用前にあらかじめ計算しておき，その他の推薦アルゴリズムに関しては合口の運用中に動的に計算した．合口はユーザが発表ページ内の文字列をマウスカーソルで選択すると，各アルゴリズムにつき最大 2 つのページを推薦し，表示はランダムに並べた．ユーザにはこれらの推薦アルゴリズムや表示方法については知らせていない．ユーザがどの推薦アルゴリズムを選択したのかについては，ユーザが合口によって推薦された他の発表ページへのリンクをクリックしたときに，そのリンクを推薦するために用いたアルゴリズムを選択したとする．

5.3 運 用 結 果

合口が対象とした Web ページは論文のタイトルや発表者，概要を含む発表ページで，全部で 276 ページあった．合口は学会の開催前から運用しており，分析対象としたデータは，2006 年 5 月 22 日（月）〜6 月 9 日（金）までの運用によって得られたデータである．運用の結果，開発者を除く 40 人のユーザが 1 回は発表ページ内の文字列をマウスカーソルで選択し，そのうち 27 人が提示された推薦リンクをクリックした．また，開発者を除く 83 人のユーザが 1 回は発表ページ上のマーキング文字列をマウスカーソルでなぞり，そ

表 4 システムが各推薦アルゴリズムによって推薦したページ数とユーザによって選択されたページ数
Table 4 Number of recommended pages based on each recommendation algorithm and selected pages by users.

推薦アルゴリズム	A	B	C	D
学会前（5/22–6/6）	27/307	9/129	7/118	10/238
学会中（6/7–6/9）	7/117	1/47	11/66	11/103

表の値は，ユーザが選択したページ数/システムが推薦したページ数

のうち 32 人が提示されたリンクをクリックした．

5.3.1 マーキングされた文字列内の語をタグと見なした場合

本項では，ユーザがマーキングを付与した文字列内の語をタグと見なした場合，マーキングによるタグ付けがソーシャルタギングとして機能するのかを調べた．ここでは，ユーザが発表ページ内の文字列をマウスカーソルで選択したときに，システムが推薦した他の発表ページのうち，どの推薦アルゴリズムによる推薦を選択したのかについて調査した．ユーザがマウスカーソルで発表ページ内の文字列を選択したときにシステムが提示した推薦リンクをクリックしたことがあるユーザのうち，学会前に使用していたのは 20 人で，学会前だけ使用していたユーザは 14 人だった．一方で，学会中に使用していたのは 13 人で，学会中だけ使用していたユーザは 7 人だった．このように，学会前と学会中とでシステムの利用者が異なるため，2 つの期間に分けて調査をした．

表 4 は，ユーザが発表ページ内の文字列をマウスカーソルで選択したときに，システムが各推薦アルゴリズムによって推薦した発表ページの数と，推薦された発表ページのうちユーザが選択した発表ページの数を示している．また，**図 10** は，各推薦アルゴリズムによって推薦された発表ページのうちユーザが選択した割合（ユーザが選択した発表ページ数/システムが推薦した発表ページ数 $\times 100$）を示している．これによると，ユーザが学会前に最も選択した推薦アルゴリズムは A で，次は推薦アルゴリズム B である．学会前にシステムを使用したユーザは，マウスカーソルで選択した文字列内の語や他の発表ページに付与されているマーキング文字列内の語を使わない推薦による発表ページを選択していた．

一方で学会中にシステムを使用したユーザは，推薦アルゴリズム C と D によって推薦された発表ページを選択していた．なかでも，推薦アルゴリズム C による推薦が推薦アルゴリズム D による推薦よりも選択されていた．これは，発表ページ内において，マーキングが付与された語を使った推薦が，文書全体の語を使った推薦よりも選択されたことになる．また，推薦アルゴリズム C による推薦は，推薦アルゴリズム A による推薦よりも選択されていることから，tfidf という統計的手法を用いた語よりもユーザがマーキングを付与した語を選択するといえる．すなわち，学会中にシステムを利用したユーザは，マーキング文字列内の語を特徴語とした Web ページを選択するということを意味している．

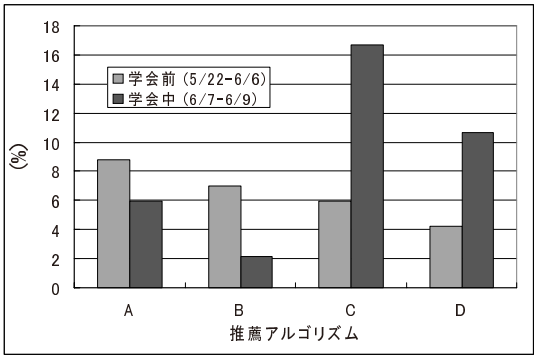


図 10 各推薦アルゴリズムによって推薦された発表ページのうちユーザが選択した割合
Fig. 10 Ratio of selected pages when users select characters in web pages.

学会前にシステムを利用したユーザは，メールによるシステム運用の告知があった次の日に，システムに推薦されたページを多く選択していたので☆，試しにシステムを利用したユーザが多かったものと思われる．学会前に選択された各推薦アルゴリズムを見ても，選択された割合に目立った差はないため，ユーザは推薦されたページの中からランダムに選択した可能性がある．

一方で，学会中にシステムを利用したユーザは，推薦アルゴリズム C と D による推薦ページを選ぶ割合が他の推薦アルゴリズムによるものより高いため，マウスカーソルで選択した文字列内の語と関連のあるページを選択することが示唆された．これは，学会中にシステムを利用したユーザは文書内で注目した語に関するページを探すという目的を持って使用したものと思われる．このような状況においては，マーキングによるタグ付けがソーシャルタギングとして機能する可能性がある．

5.3.2 マーキングが付与された文字列をタグと見なした場合

合口では，ユーザが発表ページ上に付与されている

☆ 学会前に推薦ページが選択された回数の 1 日の平均値は 4.1 回で，システム運用の告知があった次の日に推薦ページが選択された回数は 15 回だった．

表 5 システムが推薦リンクや足跡リンクとして推薦したページ数とユーザが選択したページ数
Table 5 Number of recommended pages by recommendation links and footprint links and selected pages by users.

推薦アルゴリズム	推薦リンク	足跡リンク
学会前（5/22–6/6）	20/2825	53/682
学会中（6/7–6/9）	8/2468	17/622

表の値は，ユーザが選択したページ数/システムが推薦したページ数

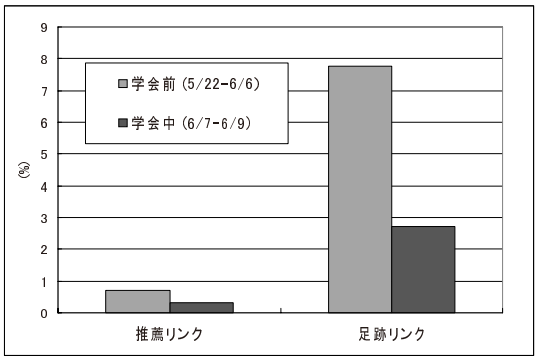


図 11 推薦リンクや足跡リンクとして推薦された発表ページのうちユーザが選択した割合
Fig. 11 Ratio of selected pages when users placed cursors over a link anchor.

マーキング文字列をマウスカーソルでなぞると，足跡リンクと推薦リンクが書かれた小窓を表示した（図 9）．本項では，ユーザがマーキングを付与した文字列をタグと見なした場合に，ソーシャルタギングとして機能するのかを調べる．足跡リンクはユーザがマーキング文字列から他の発表ページへ張ったリンクなので，マーキング文字列と直接関係のあるページと見なすことができる．一方で推薦リンクは 5.2 節で示したように，マーキングが付与された文字列内の語を利用した推薦を行っているので，マーキング文字列そのものと直接関係がある発表ページが提示されるわけではない．したがって，マーキングが付与された文字列をタグと見なしたときのリンクとそうでないときのリンクの選択回数を比較することによって，マーキングが付与された文字列をタグと見なした場合にソーシャルタギングとして機能するかどうかを調べる．

表 5 は合口が推薦リンクや足跡リンクとして推薦したページ数と，推薦された発表ページのうちユーザが選択したページ数を示している．推薦リンクや足跡リンクとして推薦された発表ページのうちユーザが選択した割合を示した**図 11** によると，学会前と学会中の両方の期間においてユーザは推薦リンクよりも足跡リンクによって提示されたページを選択していた．合口では図 9 に示すように，足跡リンクの意味合いをユーザに明示化していたことから，ユーザは意図して推薦リンクよりも足跡リンクを選んだ可能性がある．この場合，マーキング文字列がソーシャルタギングとして機能していたといえる．一方で，システムとしての利便性を向上させるために，足跡リンクをつねにリストの最上段に表示したため，ユーザは単に上に表示されていた足跡リンクを選択した可能性も考えられる．したがって，ユーザがどういう意図で足跡リンクのほうを選択したのかは定かではないが，マーキングが付与された文字列がタグとして機能することが示唆された．

6. まとめと今後の課題

本研究では，マーキングを用いたソーシャルタギングの有効性を検証するために，JSAI で運用したマーキング共有システムで得られたデータを分析した．JSAI2005 におけるイロノミーの運用で得られたデータの分析より，全ユーザで見ると色にかかわらず tfidf 値の高い語に下線が付与される可能性が高いということが分かった．しかし，マーキングされた文字列内の語の中には，tfidf 値の低い語も含まれるので，一概に tfidf 値の高い語が好まれるとはいえない．そこで我々は，JSAI2006 において，tfidf を用いたページ間類似度やマーキングされた文字列内の語を使ったページ推薦を行う合口を運用して，マーキングによるタグ付けが機能するのかについて調べた．その結果，ユーザは学会中において，tfidf によるページ間類似度によるページ推薦よりも，他のページに付与されているマーキング文字列内の語を使ったページ推薦を選択することが示唆された．また，ユーザはソーシャルブックマークのタグに相当するマーキング文字列が付与された Web ページを選択することも示唆された．しかし，マーキング文字列と同じ文字列が異なる Web ページに出現するとは限らない．これらの結果より，マーキング文字列内の語をタグと見なす手法が機能する可能性が見出された．

本研究では人工知能学会全国大会で発表された論文の概要と同大会の参加者を対象に実験を行った．したがって，比較的同じ興味を持ったユーザ集団であり，対象とした情報も均質であった．今後は，異なるコミュニティにおいての実験や，マーキングを付与する対象ページを Web 全体にした実験を行っていく必要があろう．そのうえで，マーキングによって生成されたタグからのオントロジー抽出へ進めていく必要がある．

謝辞 本研究の一部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

1) Berners-Lee, T., Hendler, J. and Lassila, O.: The Semantic Web, *Scientific American* (May 2001).
2) Brickley, D. and Guha, R.V. (Eds.): RDF Vocabulary Description Language 1.0: RDF Schema (2004).
3) McGuinness, D.L. and van Harmelen, F.: OWL Web ontology language overview, W3C Recommendation (2004).
4) Mika, P.: Ontologies are us: A Unified Model of Social Networks and Semantics, *Proc. 4th International Semantic Web Conference* (*ISWC2005*) (2005).
5) Wu, X., Zhang, L. and Yu, Y.: Exploring social annotations for the semantic web, *WWW 2006*, pp.417–426 (2006).
6) Cimiano, P., Handschuh, S. and Staab, S.: Towards the self-annotating web, *WWW 2004*, pp.462–471 (2004).
7) Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H. and Shadbolt, N.: Automatic Ontology-Based Knowledge Extraction from Web Documents, *IEEE Intelligent Systems*, Vol.18, No.1, pp.14–21 (2003).
8) Specia, L. and Motta, E.: Integrating Folksonomies with the Semantic Web, *Proc. European Semantic Web Conference* (*ESWC 2007*), Innsbruck, Austria, Springer (2007).
9) Van Damme, C., Hepp, M. and Siorpaes, K.: FolksOntology: An Integrated Approach for Turning Folksonomies into Ontologies, *Bridging the Gep between Semantic Web and Web 2.0* (*SemNet 2007*), pp.57–70 (2007).
10) Handschuh, S., Staab, S. and Studer, R.: Leveraging metadata creation for the Semantic Web with CREAM, KIM2003 advances in artificial intelligence, *Proc. Annual German Conference on AI*, September 2003 (2003).
11) Heflin, J. and Hendler, J.: A portrait of the Semantic Web in action, *IEEE Intell. Syst.*, Vol.16, No.2, pp.54–59 (2001).
12) Bechhofer, S. and Goble, C.: Towards annotation using DAML+ OIL, *Proc. Workshop on Semantic Markup and Annotation at 1st International Conference on Knowledge Capture* (*K-CAP 2001*), Victoria, B.C., Canada (2001).
13) Golder, S.A. and Huberman, B.A.: Usage patterns of collaborative tagging systems, *Journal of Information Science*, Vol.32, Issue 2 (Apr. 2006).
14) Mathes, A.: Folksonomies – cooperative classification and communication through shared metadata, Computer Mediated Communication, LIS590CMC (Doctoral Seminar), Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (Dec. 2004).
15) 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，第 20 回人工知能学会全国大会（JSAI2006）論文集 (2006).
16) 齋藤　孝：三色ボールペン情報活用術，角川書店，ISBN:4047041351 (2003).
17) 松本裕治ほか：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).
18) 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).
19) 松岡有希，坂本竜基，伊藤禎宣，武田英明，小暮潔：選択文字列を用いた Web ページ推薦システムでのユーザ参加型リンクアンカ付与機能の実証実験による評価，第 21 回人工知能学会全国大会（JSAI2007）論文集 (2007).
20) Salton, G.: Developments in automatic text retrieval, *Science*, Vol.253, pp.974–980 (1991).
21) Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J.: GroupLens: An Open Architecture for Collaborative Filtering of Netnews, *Proc. ACM Conf. on Computer Supported Cooperative Work*, Chapel Hill, North Carolina, U.S.A, pp.175–186 (Oct. 1994).

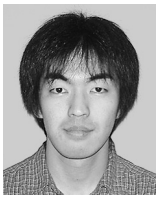
(平成 19 年 4 月 6 日受付)
(平成 19 年 9 月 3 日採録)

松岡　有希

2005 年奈良女子大学大学院人間文化研究科情報科学専攻修士課程修了．現在，総合研究大学院大学複合科学研究科情報学専攻博士後期課程に在籍．知識共有，セマンティック Web に関する研究に従事．人工知能学会学生会員．

坂本　竜基（正会員）

2003 年北陸先端科学技術大学院大学知識科学研究科博士課程修了．同年 ATR 知能ロボティクス研究所研究員．現在，ATR 知識科学研究所研究員．2005 年より奈良女子大学非常勤講師．博士（知識科学）．CSCW および HCI の研究開発に従事．ACM，日本創造学会各会員．



伊藤　禎宣（正会員）

2003 年北陸先端科学技術大学院大学知識科学研究科博士後期課程修了．博士（知識科学）．同年 ATR メディア情報科学研究所研究員．2006 年 NICT ユニバーサルメディア研究センター短時間研究員，ATR 知識科学研究所客員研究員，東京農工大学大学院工学府特任講師．2007 年より同特任准教授．HCI，CSCW に興味を持つ．

大向　一輝（正会員）

2000 年同志社大学工学部知識工学科卒業．2002 年同大学院工学研究科知識工学専攻博士前期課程修了．2005 年総合研究大学院大学複合科学研究科博士後期課程修了．博士（情報学）．同年国立情報学研究所助手．2006 年総合研究大学院大学助手（併任）．2007 年より同助教．2003 年度情報処理振興機構未踏ソフトウェア創造事業スーパークリエータ．セマンティックウェブ，ソーシャルネットワークを用いた知識共有の研究に従事．人工知能学会，電子情報通信学会各会員．

武田　英明（正会員）

1986 年 3 月東京大学工学部卒業．1988 年 3 月東京大学大学院修士課程，1991 年 3 月博士課程修了．工学博士．ノルウエー工科大学，奈良先端科学技術大学院大学を経て，2000 年 4 月国立情報学研究所助教授，2003 年 5 月同教授．2006 年 4 月同学術コンテンツサービス研究開発センター長（併任）．東京大学人工物工学研究センター客員教授，大阪大学 RISS 特任教授，ATR 知識科学研究所客員研究員（兼務）．知識共有，設計学等の研究に従事．人工知能学会，AAAI 各会員．

小暮　　潔（正会員）

1981 年慶應義塾大学大学院工学研究科電気工学専攻修士課程修了．同年日本電信電話公社に入社．現在，ATR 知識科学研究所所長．博士（工学）．自然言語処理，エージェント，ロボット，知的環境等の研究に従事．電子情報通信学会，人工知能学会，言語処理学会，日本認知科学会，日本音響学会各会員．