ソーシャルタグの組合せに基づく希少なwebコンテンツの推薦

多田　亮平† 湯本　高行†† 新居　学†† 高橋　豊†† 角谷　和俊†††

† 兵庫県立大学工学部〒671-2280 兵庫県姫路市書写 2167
†† 兵庫県立大学大学院工学研究科〒671-2280 兵庫県姫路市書写 2167
††† 兵庫県立大学環境人間学部〒670-0092 兵庫県姫路市新在家本町 1-1-12
E-mail: †eo07e058@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,
†††sumiya@shse.u-hyogo.ac.jp

あらまし 近年，検索技術の向上により目的の情報の取得が容易になっているが，情報の発見性の点では不十分である．本研究では希少なwebコンテンツを推薦し，ユーザに情報の発見を促すことを目的とする．本稿ではソーシャルブックマークにおいてwebコンテンツに付与されるタグに注目する．与えられたタグに対して，そのタグを持つコンテンツに付与された他のタグを共起タグとして取り出す．与えられたタグと各共起タグの組合わせを同時に付与されたコンテンツが少ないものを意外，それを支持するユーザが多いものを有用と定義する．有用性が一定以上で，意外な順にソートしたタグの組合せを提示し，それが付与されたwebコンテンツを推薦する．意外な順と，その逆順を比較することで，本手法により意外なタグの組合せを取得しやすいことが分かった．

キーワード ソーシャルブック，情報推薦，希少情報

Recommendation of Rare Web Contents based on Combination of Social Tags

Ryohei TADA†, Takayuki YUMOTO††, Manabu NII††, Yutaka TAKAHASHI††, and Kazutoshi SUMIYA†††

† School of Engineering, University of Hyogo 2167 Shosha，Himeji，Hyogo，671-2280，Japan
†† Graduate School of Engineering, University of Hyogo 2167 Shosha，Himeji，Hyogo，671-2280，Japan
††† School of Human Science and Environment, University of Hyogo.1-1-12 Shinzaike-honcho, Himeji,Hyogo 670-0092, Japan
E-mail: †eo07e058@steng.u-hyogo.ac.jp, ††{yumoto,nii,takahasi}@eng.u-hyogo.ac.jp,
†††sumiya@shse.u-hyogo.ac.jp

1. はじめに

近年，インターネットを用いた情報の検索技術の向上により，目的の情報の取得が容易になった．例えば検索エンジンでは，クエリの推薦機能が追加されたことで，ユーザが入力したクエリに対して一般的に他のユーザが同時に入力する他のクエリが提示される．しかし，このように検索エンジンを用いた場合，推薦されるコンテンツは一般的な情報に偏ってしまい，ユーザにとって既知の情報が推薦される可能性が高い．そのため，ユーザが知らない情報を欲する場合には，検索エンジンを用いた情報の検索は不十分であると考えられる．またユーザが一部の人しか知らない価値のある情報を探している場合も，あまり認知されていない情報は検索結果の上位には上がりにくいため，検索エンジンを用いた検索ではユーザにとって負担は大きいと考えられる．そこで本研究ではソーシャルブックマークサービス (以下 SBM サービス) で用いられるタグ (ソーシャルタグ) に注目し，ユーザにとって未知である可能性が高い，希少な情報を推薦する手法を提案する．SBM サービス内において，ユーザはタグという関連キーワードやカテゴリを付与することでwebコンテンツを分類する．これにより各webコンテンツはタグという要素を持っていると考えることで，webコンテンツに関連する情報をこのタグから得ることができる．一般的でない組合せのタグを持ったwebコンテンツは意外な情報であると仮定し，その中でも有用であるwebコンテンツを判別す

ることで希少な情報の推薦を行う．

2. 関連研究

ソーシャルタグに基づいたコンテンツの推薦の関連研究として協調フィルタリングを用いた研究が存在する．例えば佐々木ら [1] は，入力ユーザの嗜好に基づいた情報を推薦するというシステムを提案している．入力ユーザのブックマーク集合から似た趣味のユーザを推定し，そのブックマーク集合から推薦先のユーザが登録していない web コンテンツを推薦する．入力ユーザの嗜好に似た別のユーザのブックマーク情報を推薦するため，ユーザの嗜好に基づいた推薦としては精度も高く有用である．しかしこの手法ではユーザは SBM サービスに登録し，いくつかタグ付きでブックマークを登録する手間が発生する．また協調フィルタリングを用いて高い精度で嗜好の似た情報を推薦するため，推薦情報に多様性が失われユーザにとって既知の情報を推薦する可能性が高い．本研究で提案する手法では，一般的に SBM サービスユーザ全体にとって検索語に関連する希少な web コンテンツを推薦することを目的としている．そのため，ユーザにとって未知の情報を推薦するという点で推薦する情報や目的の違いがある．また SBM サービスで既にあるデータを使用するため，ユーザがデータを入力する負担を減らして推薦することができる．

本研究と同様に，情報の発見性について着目した研究として，清水ら [2] の研究がある．この研究では情報の発見性に加えて，情報がユーザの嗜好に合っているかという要素も同時に扱っており，ユーザが「知らない」かつ「好みである」Novelty という指標に対して最適なアルゴリズムを考案している．また推薦するアイテムに関しては楽曲データとなっており，本研究のタグや web コンテンツではない．情報の発見性の点では，ユーザ自身が知らないアイテムを高い確率で推薦することが可能となっている．嗜好性も考慮した Novelty は推薦上位 10 件に対して 4 割程度となっており，新しい曲を望むユーザにとって非常に有用である．しかし，こちらも協調フィルタリングを用いて推薦を行うため，高い精度を得るには「既知か未知か」,「好みであるか」というユーザが入力するテストデータを多く必要とする．そのため，システムを利用し始めるユーザにとって，負担は大きくなるといえる．本研究は SBM サービスを利用しているユーザの傾向を利用して情報を推薦するため，ユーザ自身がデータを入力するという負担はない．

3. 希少情報の推定手法

3.1 希少情報の定義

あるカテゴリやトピックを考えたとき，そこから得られる希少な情報は，以下の 2 つの場合が考えられる．1 つ目は，特定のトピックに対して，それを細分化したサブトピックの中でも特にマイナーなサブトピックの情報がある．例えば，トピックをプログラミング言語である「java」とした場合,「java」に対してマイナーな「java のライブラリ」というトピックの情報がこれにあたる．2 つ目は，特定のカテゴリに対して，同時に分類する例が少ないカテゴリと組み合せた情報にあたる．例えばカテゴリを「java」とした場合,「java」と「Firefox」というカテゴリを組み合わせたマイナーな情報がこれにあたる．これは 1 つ目に説明した希少情報とは違い，ウェブブラウザである「Firefox」は，プログラミング言語である「java」を細分化したカテゴリではない．逆に考えたときも同様に,「java」は「Firefox」を細分化したカテゴリではない．本研究ではこの 2 つ目の，同時に分類する例が少ないカテゴリと組み合せた情報を「希少情報」として定義する

以下 3.2 章で SBM サービスのタグという機能に注目し，入力に対する意外なタグの組合わせについて述べる．また 3.3 ではどれだけマイナーであるかを測る尺度である意外度，3.4 ではどれだけ有用であるかを測る尺度である支持度について述べる．

3.2 意外なタグの組合せの定義

ユーザがコンテンツに対してタグを付けてブックマーク登録する際，どのようにタグが使われているかを Golder ら [3] が調査した，その例を以下に示す．

トピックタグ…ブックマークした Web ページのトピックを表す (例：“料理”，“Sports”)

種類タグ…ブックマークした Web ページの種類を表す (例：“本”，“blog”)

著者タグ…ブックマーク Web ページの著者，サイトを表す

詳細タグ…共起しているタグが指す範囲を限定するタグ (例：(.net 等バージョン情報として)“3.5”)

主観タグ…ブックマークした Web ページの評価や感想を表す (例：“fanny”，“これはすごい”)

関与タグ…“my”で始まるタグで，Web ページにどのような関与をしているかを表す (例：“mycomment”，“mystuff”)

タスクタグ…ユーザが Web ページをどう扱うかを表す (例：“後で読む”,“jobsearch”)

上の分類の中でも，特にトピックタグ，種類タグについて SBM サービスではよく用いられている．このことから，ユーザはブックマークしたコンテンツをタグによって「カテゴリ」ごとに分類していると考えられる．ここで，カテゴリの組合せが少ない例を「意外なカテゴリの組合せ」と考える．そのため「意外なタグの組合せ」を取得することで，希少情報が得られると考えられる．

本研究での入力は，SBM サービスに既に登録されているタグを用い，そのタグを「クエリタグ」と呼ぶ．またブックマーク登録をする際に，クエリタグと同時に使用された他のタグを「共起タグ」と呼ぶ．クエリタグと共起タグ，クエリタグが付与されたコンテンツ集合と共起タグが付与されたコンテンツ集合の関係の例を図 1 に示す．共起タグ A，B，C，D はクエリタグに対して以下のような関係にあると考える．

タグ A…クエリタグに対してよく知られている組合せのタグ

タグ B…クエリタグの上位語に当たる組合せのタグ

タグ C…クエリタグに対して意外な組合せのタグ，もしくは関与タグやタスクタグ

タグ D…全く関連のないタグ

本研究ではクエリタグを付与されたコンテンツの集合を推薦

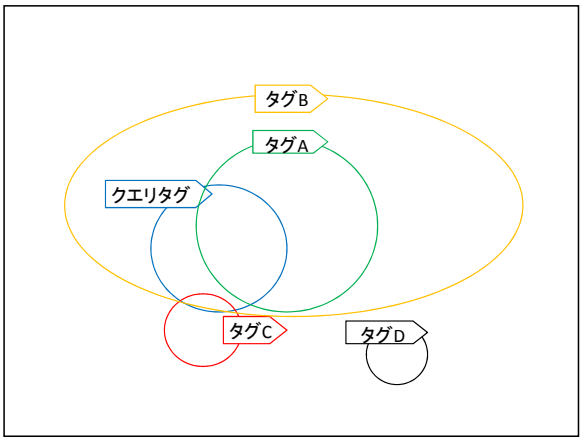

図 1　クエリタグに対する他のタグの関係

の候補とする．このうちタグ C と，クエリタグの共起するコンテンツ集合を推薦する．クエリタグと，これらのタグとの関係を例を踏まえて説明する．

まずタグ A が付与されたコンテンツ集合はクエリタグのコンテンツ集合と大きく共起している．これは SBM サービスユーザ全体からクエリタグとタグ A は非常に関連していると判断されているとみなし，一般的なタグの組合せを持つコンテンツと考える．例えばクエリタグを「java」とした場合，java の統合開発環境である「netbeans」のようなキーワードや，java を用いて解析することが多い「xml」などがタグ A にあたる．

また，タグ B が付与されたコンテンツ集合はクエリタグのコンテンツ集合を包含している．これは SBM ユーザ全体から，クエリタグの上位語によってコンテンツを分類するためにタグ登録されていると判断できる．先程の例と同様にクエリタグを「java」とした場合，タグ B は「programming」といったタグにあたる．

次にタグ C について説明する．クエリタグのコンテンツ集合とタグ C のコンテンツ集合の共起は小さい．これはクエリタグとタグ C の関連は少数の SBM ユーザにしか認知されていないため，一般的でない意外なタグの組合せである場合と，ユーザの入力ミスや関与タグ，タスクタグのように推薦に適さないノイズとなるような組合せである場合が考えられる．例としてクエリタグが「java」とする．ことのきタグ C が「thunderbird」や「tex」となる場合，メーラーである thunderbird 用に java で開発されたスパムフィルタを説明するコンテンツや，java で開発された tex を作成するアプリケーションのコンテンツがあるため，ノイズではない意外なタグの組合せだといえる．しかしタグ C が「*あとで読む」となる場合，あるユーザがあとで自分で読むために一時的に登録する，独自ルールに基づくタグとなり意外であるとは言えない．

最後にタグ D についての説明をする．付与されたクエリタグのコンテンツ集合と一切共起しない，関連のない情報と判断されるコンテンツ集合がある．これはクエリタグと関連がないため，一般的でないタグの組合せではあるが，推薦には適さない情報として推薦の範囲から排除する．例えば同様にクエリタグを「java」とした場合，中性子を指す「neutron」のような物理に関する情報はクエリタグに関係がないため，推薦に適さない．

本稿ではこのうちクエリタグとタグ C を意外なタグの組合せと考え，その中でも特に有用と考えられる組合せのタグと，その組合せを付与されたコンテンツの集合を推薦する．

3.3　タグの組合せに基づく意外性の判定

本研究ではクエリタグと共起が小さいタグを持つコンテンツを意外な情報と考える．このときクエリタグが付与されたコンテンツ集合と，共起タグが付与されたコンテンツ集合の大きさを考慮すると，双方の集合から考えても共起が小さくなくてはいけない．そのため，共起の度合いを測る指標としてシンプソン係数を指標として用いる．本研究ではシンプソン係数 $Simpson$ は，クエリタグと共起タグで共起したコンテンツ数を，2 つのコンテンツ集合のうち小さい集合の要素数で割ったものとして扱う．タグ t_1，t_2 が付与されたコンテンツ集合をそれぞれ C_{t_1}，C_{t_2} として (1) 式のように算出する．

$$Simpson(t_1, t_2) = \frac{|C_{t_1} \cap C_{t_2}|}{min(|C_{t_1}|, |C_{t_2}|)} \qquad (1)$$

ここで $min(|C_{t_1}|, |C_{t_2}|)$ は，タグ t_1，t_2 が付与されたコンテンツ集合のうち，小さいコンテンツ集合の要素数を表す．このシンプソン係数が小さいほど，共起が一般的でない意外なタグの組合せであると仮定する．そのためクエリタグ，共起タグをそれぞれ t_q，t_c とし，これらから算出される意外度 $Unexp$ を (2) 式のように定義する．

$$Unexp(t_q, t_c) = 1 - Simpson(t_q, t_c) \qquad (2)$$

3.4　有用性の判定

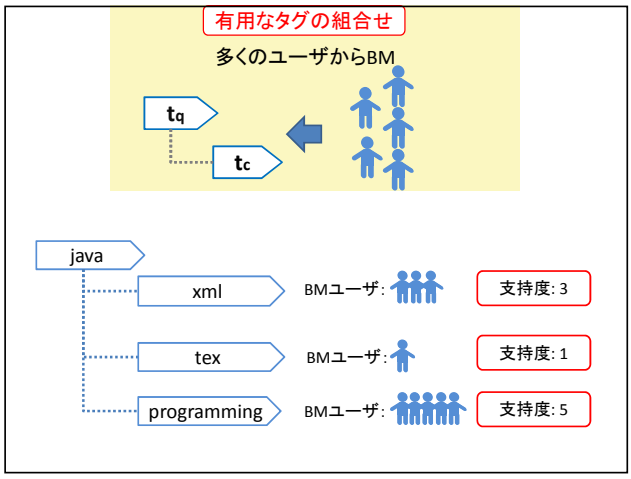

図 2　支持度の算出方法

3.2 で述べたように，クエリタグと共起の小さいタグとの組合せが，一般的でない意外なタグの組合せか，ノイズかはタグ同士の共起を測定するだけでは不明である．よってタグの組合せの有用性を測る指標が必要である．その手法を以下に示す．有用性を測るためのイメージ図を図 2 に示す．図 2 上部のようにクエリタグ “t_q” と共起したタグ “t_c” の組合せがあると考える．多数のユーザが “t_q” と “t_c” を同時に使用している場合，このタグの組合せはその多数のユーザにとって支持されているた

め有用であると考えられる．またユーザのタグ登録のミスや，独自のルールでタグを付与しているような有用ではないタグの組合せについては，そのタグの組合せを用いてブックマークしているユーザ数が少ないと考えられる．以上の二点から，クエリタグ t_q に対してそれぞれの共起タグ t_c との組合せを支持しているユーザ数を支持度とし，(3) 式のように定義する．ただしここで，U_T は $t \in T$ の全てのタグ同時に使用しているユーザの集合として示している．

$$Popularity(t_q, t_c) = |U_{t_q,t_c}| \tag{3}$$

3.5 希少なタグの組合せの推薦方法

3.3 と 3.4 の手法で算出した意外度と支持度を組み合せてユーザへ希少なタグの組合せを推薦する．ユーザへ提示する推薦の形式を図 3 に示す．入力したクエリタグに対して共起タグを取得し，各共起タグの意外度と支持度を算出する．それぞれのタグの組合せについて，以下の処理を行い，出力する．

（1） **意外度による降順ソート**: 意外度が大きい順にソートすることで，カテゴリの組合せが意外であるコンテンツ集合を推薦する．

（2） **支持度によるフィルタリング**: 支持度が平均値以上のタグの組合せを出力することで，ある程度のユーザに支持されるタグの組合せを出力する．

また，全体を意外度によってソートしているため，支持度が平均値以上のタグの組合せだけを出力しても，推薦上位には意外なタグの組合せが出力されると考えられる．このように並び替えることで有用かつ意外なタグの組合せを出力する．また，それぞれのタグの組合せを持つコンテンツ集合についてはブックマーク数が多い順に並び替える．これによって多くのユーザに支持されるコンテンツを推薦する．これらを図 3 のようにツリー形式で表示することで，ユーザへ視覚的に推薦する．

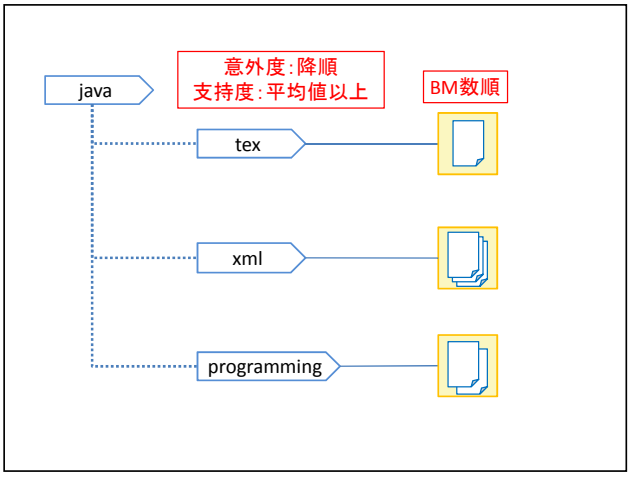

図 3 推 薦 形 式

4. 実　　験

4.1 実 験 方 法

提案手法による推薦が有効であるかを確認するために，SBM サービス (livedoor クリップ [4]) の提供によるデータベースを用いて，表 1 のタグを入力とし，の組合せに基づくコンテンツを推薦する．

表 1 入力に用いるクエリタグ一覧

google	ゲーム	windows	firefox	twitter
music	経済	アニメ	アルバイト	ダイエット

表 2 被験者へ提示する組合せ

	フィルタリング	
	あり	なし
意外度降順	a	b
意外度昇順	c	d

各手法が有効であるかを判定するために，意外度によって「意外なタグの組合せが推薦されているか」，支持度によって「有用なタグの組合せが推薦されているか」をそれぞれ確認する．まず各クエリタグについて，以下で説明する各方法での上位 10 個のタグの組合せをそれぞれ 7 人の被験者へ評価してもらう．被験者に共起タグ集合を表 2 の a〜d の 4 種類の場合の提示をする．表 2 において，意外度降順は意外度が大きい順にソートしたもの，意外度昇順は意外度の小さい順にソートしたものを表している．また，フィルタリングは「支持度が平均値以上の組合せだけを推薦する」というフィルタリングである．この中では a が提案手法の提示方法となる．被験者にはクエリタグに対して，各共起タグが,「意外か，一般的か，ノイズなどの意味が分からないタグか」という 3 段階の評価をしてもらう．次に「その組合せがコンテンツと対応しているか，対応していないか」の 2 段階評価，タグの組合せをもつコンテンツの中で，最もブックマーク数が多かったコンテンツについてその情報が「既知か，未知か」の評価を得る．

これらの提示方法で得られるタグの和集合を取り，タグとコンテンツ集合のセットをランダムな順番で提示して評価を得ることで，被験者の先入観を排除して希少性についての評価を得る．

4.2 実 験 結 果

まず，被験者 7 人に評価実験を行ってもらい，表 2 における a，b，c での提示方法で得られる各評価結果を比較することで，手法の評価を行う．a，b，c の各提示方法で得られる「意外かつコンテンツと対応の取れたタグ」の個数を表 3，4 ，5 に示す．次に，実際にそれぞれの提示方法で得られる共起タグの例を表 8，9，10，11 に示す．これらの表中の共起タグは，それぞれ左から順に推薦順となっている．

意外度の有効性を調べるために，a，c での提示方法で得られる各評価結果を比較する．列を被験者 A〜G，行を各クエリタグとし，表 6 に示す．各要素には表 2 の a，c それぞれでの提示方法で得られるタグから,「意外かつコンテンツと対応の取れたタグ」と評価された個数を算出し，“c で得られた個数”→“a で得られた個数”として示している．矢印の先の数が大きい場合，意外な順に並び替えることで意外なタグの推薦数が増加

表 3　(a) 意外度降順フィルタありの提示結果

クエリタグ	A	B	C	D	E	F	G	クエリ平均
google	0	1	5	4	3	5	6	3.4
ゲーム	3	5	4	5	6	7	6	5.1
windows	3	1	4	0	1	3	4	2.2
firefox	2	5	6	3	4	5	4	4.1
twitter	2	2	7	2	1	6	5	3.5
music	5	4	8	6	5	4	4	5.1
経済	3	6	5	5	5	1	6	4.4
アニメ	4	5	5	4	1	3	5	3.8
アルバイト	2	1	3	3	0	2	0	1.5
ダイエット	1	1	1	1	3	1	3	1.5
ユーザ平均	2.5	3.1	4.8	3.3	2.9	3.7	4.3	3.5

表 4　(b) 意外度昇順フィルタなしの提示結果

クエリタグ	A	B	C	D	E	F	G	クエリ平均
google	0	1	4	3	2	3	6	2.7
ゲーム	1	2	3	1	2	2	2	1.8
windows	4	4	3	4	1	4	5	3.5
firefox	2	4	8	4	5	6	7	5.1
twitter	0	4	3	3	5	3	3	3.0
music	2	6	3	4	5	3	4	3.8
経済	4	4	5	5	5	4	4	4.4
アニメ	7	3	5	5	3	5	5	4.7
アルバイト	2	0	1	1	0	2	0	0.8
ダイエット	4	2	2	3	5	4	4	3.4
ユーザ平均	2.6	3	3.7	3.3	3.3	3.6	4	3.3

表 5　(c) 意外度昇順フィルタありの提示結果

クエリタグ	A	B	C	D	E	F	G	クエリ平均
google	2	2	8	8	4	8	2	4.8
ゲーム	0	0	0	1	0	0	1	0.2
windows	4	0	4	4	2	4	2	2.8
firefox	4	0	1	3	1	4	2	2.1
twitter	4	1	6	3	4	6	2	3.7
music	4	0	2	3	2	5	0	2.2
経済	0	0	0	2	0	0	0	0.2
アニメ	4	1	1	1	1	2	1	1.5
アルバイト	0	0	0	0	0	0	0	0
ダイエット	1	0	0	1	0	0	0	0.2
ユーザ平均	2.3	0.4	2.2	2.6	1.4	2.9	1.0	1.8

したと考える．各クエリタグに対して,「意外かつコンテンツと対応の取れたタグ」の推薦数が増加した場合の数を被験者数で割ったものを「クエリごとの増加率」，各ユーザに対してクエリタグの数で割ったものを「ユーザ毎の増加率」とする.「クエリごとの増加率」は各クエリでどれだけ意外なタグが得やすくなったか,「ユーザ毎の増加率」は各ユーザに対してどれだけ意外なタグを提示しやすくなったかを表す.

表 6 のクエリ毎ごとの増加率から，提示方法を c から a にすることによって，多くのクエリタグに対して意外なタグの推薦個数が増加した．このことから意外度によって意外なタグが得やすくなったといえる．特にクエリタグが「ゲーム」,「経済」の場合，全てのユーザに対して推薦する意外な組合せのタグが増加した．次に，実際にどの程度の意外な組合せのタグが推薦できるかを，表 3，5 を比較して調べる．表 3，5 の右下の “意外かつ対応が取れているタグ”の平均推薦個数は，提示方法を c から a にすることで，1.8 個から 3.5 個へ増加することがわかる．表 6 で被験者全員に対して有効だったクエリタグである「ゲーム」に関して，c では意外なタグは平均で 0.2 個しか得られなかったが，a では平均で 5.1 個得られ，大きく増加したことがわかる．同様に「経済」に関しては，0.2 個だったものが 4.4 個と，得られる “意外かつ対応が取れているタグ”の個数は大きく増加した．実際にそれらのクエリタグで得られる共起タグは，表 8 から,「ゲーム」の a の提示方法では「firefox」や「金融」などのタグが得られる.「経済」の a の提示方法では「iphone」,「music」,「ニコニコ動画」などのタグが得られ，意外なタグが得やすくなったといえる．

しかし，手法が有効でないクエリタグもあった．表 6 から，クエリタグが「google」,「windows」の場合，クエリ毎の増加率が非常に小さい．これはその二つのクエリタグの場合，提示方法を c から a にすることで，“意外かつ対応が取れているタグ”の推薦個数が減る被験者が多いことを表している．またクエリタグが「google」の場合，表 3 の c の提示方法で得られる “意外かつ対応が取れているタグ”が 4.8 個となったが，表 5 の a の提示方法にすると 3.4 個と少なくなった．同様にクエリタグが「windows」の場合，2.8 個から 2.2 個に減少しており，これらのクエリタグでは定義した意外度が有効に機能していないことがわかる．

次に，ノイズを除去するための支持度が有用かどうかを確認する．表 6 と同様に，表 2 の提案手法である意外度降順フィルタあり (a) と，意外度降順フィルタなし (b) の提示方法での比較を表 7 に示す．表 7 のクエリごとの増加率から，提示方法を b から a にすることで，意外なタグの推薦個数が増加した例は「google」,「ゲーム」,「twitter」,「music」だけとなった．他のクエリタグについては，フィルタリングの効果が大きくは得られなかったといえる．実際にどの程度の意外な組合せのタグが推薦できるかを，表 3，表 4 を比較して確認する．表 3，表 4 の右下の “意外かつ対応が取れているタグ”の平均推薦個数は，提示方法を b から a にすることで，3.3 個から 3.5 個に僅かに増加した．表 7 から被験者全員に対して有効だとわかる「ゲーム」に関しては，b では 1.8 個しか得られなかったが，a にすることで 5.1 個まで増加した．しかし，他のクエリタグでは大きな効果は得られなかった．

実際に得られるタグの例を，表 8，9 の推薦で比較する.「支持度が平均値以上」というフィルタリングを適応することで，表 9 の「ゲーム」で提示される「[」のような登録のミスは排除している．また,「firefox」では他のタグが多くの人数に支持されているため,「@ldr」のようなタスクタグは排除できた.「windows」では「考え方」のような抽象的なタグが取得されたが，フィルタリングを行うことで排除できることがわかる．表 11 で多くみられる，使用しているユーザが明らかに少ないと思われるようなタグも，フィルタリングを行った表 10 では

表 6　ユーザ，クエリごとの手法の有効性 (c → a)

クエリタグ	A	B	C	D	E	F	G	クエリごとの増加率
google	2 → 0	2 → 1	8 → 5	8 → 4	4 → 3	8 → 5	2 → 6	0.14
ゲーム	0 → 3	0 → 5	0 → 4	1 → 5	0 → 6	0 → 7	1 → 6	1.00
windows	4 → 3	0 → 1	4 → 4	4 → 0	2 → 1	4 → 3	2 → 4	0.29
firefox	4 → 2	0 → 5	1 → 6	3 → 3	1 → 4	4 → 5	2 → 4	0.71
twitter	4 → 2	1 → 2	6 → 7	3 → 2	4 → 1	6 → 6	2 → 5	0.43
music	4 → 5	0 → 4	2 → 8	3 → 6	2 → 5	5 → 4	0 → 4	0.86
経済	0 → 3	0 → 6	0 → 5	2 → 5	0 → 5	0 → 1	0 → 6	1.00
アニメ	4 → 4	1 → 5	1 → 5	1 → 4	1 → 1	2 → 3	1 → 5	0.71
アルバイト	0 → 2	0 → 1	0 → 3	0 → 3	0 → 0	0 → 2	0 → 0	0.71
ダイエット	1 → 1	0 → 1	0 → 1	1 → 1	0 → 3	0 → 1	0 → 3	0.71
ユーザ毎の増加率	0.40	0.90	0.80	0.50	0.50	0.60	0.90	

表 7　ユーザ，クエリごとのフィルタリングの有効性 (b → a)

クエリタグ	A	B	C	D	E	F	G	クエリ毎の増加率
google	0 → 0	1 → 1	4 → 5	3 → 4	2 → 3	3 → 5	6 → 6	0.57
ゲーム	1 → 3	2 → 5	3 → 4	1 → 5	2 → 6	2 → 7	2 → 6	1.00
windows	4 → 3	4 → 1	3 → 4	4 → 0	1 → 1	4 → 3	5 → 4	0.14
firefox	2 → 2	4 → 5	8 → 6	4 → 3	5 → 4	6 → 5	7 → 4	0.14
twitter	0 → 2	4 → 2	3 → 7	3 → 2	5 → 1	3 → 6	3 → 5	0.57
music	2 → 5	6 → 4	3 → 8	4 → 6	5 → 5	3 → 4	4 → 4	0.57
経済	4 → 3	4 → 6	5 → 5	5 → 5	5 → 5	4 → 1	4 → 6	0.28
アニメ	7 → 4	3 → 5	5 → 5	5 → 5	3 → 1	5 → 5	5 → 5	0
アルバイト	2 → 2	0 → 1	1 → 3	1 → 3	0 → 0	2 → 2	0 → 0	0.42
ダイエット	4 → 1	2 → 1	2 → 1	3 → 1	5 → 3	4 → 1	4 → 3	0
ユーザ毎の増加率	0.4	0.5	0.6	0.4	0.2	0.4	0.3	

提示する推薦タグから多く排除している．しかし，どれだけのユーザに支持されているかという支持度では，表 9 で提示される “多くのユーザには支持されていない”が “意外かつ対応が取れている”タグも除去してしまう．また表 8 の「google」では「人気」のような抽象的なタグや,「@ldr」のようなタスクタグは排除できていないことがわかる．

4.3 考　　察

実際に各提示方法で得られるタグを表 8，10 に示している．「google」に関して，a の意外度降順フィルタありの提示方法では，使用した Livedoor クリップ内でよく用いられる「@ldr」,「clip」,「あとで見る」などのタグが現れており，3.2 で示したタスクタグが多く推薦上位に現れた．これにより，コンテンツ内容と対応したタグの組合せが取得できていないことが分かった．また，c の意外度昇順フィルタありの提示方法では，google が提供するアプリケーションやサービスを示す,「[chrome」,「gears」,「checkout」が多い．これらの機能の中でもあまり知られていない「checkout」などの機能が被験者にとって意外だと評価されていることで，意外度昇順によるソート中に多く意外なタグが多く現れたと考えられる．表 5,3 から,「windows」に関しては意外度昇順にソートした提示方法からは，意外なタグは平均で 2.8 個，意外度降順によるソートでは平均 2.2 個と，得られる意外なタグは少ない．表 8，10 から，意外度降順によるソートの推薦上位には「レビュー」,「購入」といったどこでも使われるような汎用的，抽象的なタグが入った．また意外度昇順にソートした提示方法には，クラウド向けの OS である Windows Azure を示す「Azure」や，システムユーティリティ等を紹介するためのサイト Sysinternals を示す「sysinternals」などが得られ，多くの被験者に意外であると評価された．

これらは，意外度はシンプソン係数によって定義し，それを逆順にソートし意外度昇順 (シンプソン係数の大きい順) として並び替えているため，3.1 で挙げた「トピックを細分化した」ような希少情報も意外度昇順の上位になることが多くなっている．そのため，被験者が「意外」だと評価する情報が多くなったと考えられる．またフィルタリングについては支持度のように，ブックマークしているユーザ数のみを対象にするようなフィルタリングでは，推薦に適さないタグが排除できない．

表 5，3 の全体での意外なタグが得られる平均数は，1.8 個から 3.5 個に増加しており，クエリの大部分に対して意外度によるソートが有効であるといえる．しかし，上位 10 件から取得できる意外なタグの個数は 3.5 個程度と多くはない．特にクエリタグが「アルバイト」,「ダイエット」の場合，得られる意外なタグの個数が非常に少ない．表 3 から，特に取得数が少ない「アルバイト」,「ダイエット」に関しては，共にリンク切れのため評価できないタグが多く取得されたことが原因にあげられる.「アルバイト」に関しては，共起タグに地名が多く取得されたことや，業者の広告によるタグ登録が多いことが原因になったといえる.「アルバイト」に関しても業者の広告によるタグ登録が多く，特定のページのブックマークが異常に多いという問

題がある．

5. おわりに

本稿では，ソーシャルタグの組合せに注目し，組合せの意外性を測る意外度，どれだけのユーザに支持されているかの支持度を用いることで，希少な web コンテンツを推薦する手法を提案し，また手法の評価実験を行った．10 件のクエリタグに対して，7 件のクエリタグにおいて「意外かつコンテンツと対応が取れたタグ」が得やすくなった．このことから，多くのクエリタグに対して一般的には得難いタグの組合せを持つ web コンテンツが得やすくなったといえる．しかし，推薦する「意外かつコンテンツと対応が取れたタグ」の個数は 1.7 個程度しか増加しない．この原因としては「*あとで読む」のような，タスクタグが推薦上位に入ること.「人気」,「レビュー」などに見られる，汎用的・抽象的な特定の意味を成さないタグが推薦上位に入ることが考えられる．また，シンプソン係数を用いて「タグの組合せの意外度」を算出しているため，ユーザが「意外」だと感じる「トピックを細分化した」ようなタグの組合せは取得できない．このような希少情報についても推薦する手法が必要だと考えられる．

しかし，被検者によって推薦できる「意外かつコンテンツと対応が取れたタグ」の個数は偏りがある．この結果から各ユーザの知識量に合わせた推薦手法が必要であるといえる．

今後の課題として，取得するタグのうちで,「*あとで読む」のような推薦に適さないようなタグや,「人気」などの抽象的なタグを排除する手法の考案，適用が必要である．また情報を推薦するユーザごとに適応させる手法が必要になる．

謝　　辞

本研究の一部は，平成 22 年度科研費基盤研究 (B)(2)「ユーザの潜在的意図を用いたレス・コンシャス情報検索基盤の構築」(課題番号：20300039) によるものです．ここに記して謝意を表すものとします．

文　　献

[1] 佐々木祥, 宮田 高道, 稲積泰宏, 小林亜樹, 酒井善則：Social-Bookmark におけるコンテンツクラスタ間の類似度を用いた web コンテンツ推薦システム, 情報処理学会論文誌データベース, Vol. 48, SIG 20(TOD36) , pp. 14–27 (2007).

[2] 清水拓也, 土方嘉徳, 西田正吾：発見性を考慮した協調フィルタリングアルゴリズム, 電子情報通信学会論文誌,Vol.J91-D,No.3,pp. 538-550(2008).

[3] S.A.Golder andB.A.Huberman: “The structure of collablorative tagging systems”,Journal of Information Science,32,2,pp.198-208(2006).

[4] “Livedoor クリップ”,http://clip.livedoor.com/.

表 8　(a) 意外度降順フィルタありでの推薦上位 10 件のタグ

クエリタグ	推薦上位の共起タグ
google	人気，@ldr，clip，科学，コミュニケーション，あとで見る，society，音楽，communication，生活
ゲーム	口コミ，firefox，security，hatena，social，便利，media，金融，science，tv
windows	レビュー，購入，clip，hatena，インターネット，book，素材，ニコニコ動画，social，デザイン
firefox	ゲーム，perl，business，ブログ，音楽，デザイン，ネット，sns，clip，サービス
twitter	toread，画像，ajax，web デザイン，生活，アニメ，ldr，デザイン，プログラミング，仕事
music	ruby，医療，library，経済，webdesign，lifehack，work，携帯電話，hatena，ui
経済	net，management，music，iphone，science，小説，ニコニコ動画，youtube，旅行，広告
アニメ	セキュリティ，marketing，video，pc，design，マーケティング，twitter，検索，メモ，neta
アルバイト	まとめ，ローン，名古屋，甘い，秋葉原，京都，大阪，パソコン，横浜，愛知
ダイエット	比較，テレビ，バッグ，本，ライブドア，無料，楽天，ブランド，ネットビジネス，裏技

表 9　(b) 意外度降順フィルタなしでの推薦上位 10 件のタグ

クエリタグ	推薦上位 10 件の共起タグ
google	評価，口コミ，通販，方法，女性，審査，netwatch，アニメ，恋愛，素材
ゲーム	評価，plaggger，toread，ldr，口コミ，lifehack，[，web デザイン，photo，ruby
windows	社会，@ldr，教育，方法，society，広告，考え方，企業，漫画，ランキング
firefox	!未読，マーケティング，@ldr，レビュー，検証，society，ニュース，科学，楽天，ゲーム
twitter	@ldr，健康，口コミ，転職，ブランド，料理，it 業界，学習，経営，レビュー
music	ldr，web デザイン，seo，画像，仕事術，恋愛，tmp，セキュリティ，情報，tech
経済	tool，toread，software，ldr，seo，mac，flash，photo，security，はてな
アニメ	tmp，hatena，plaggger，security，development，セキュリティ，image，column，photoshop，*news
アルバイト	経済，内容，あとで読む，design，ネット，work，ダイエット，海外，勉強，ソフト
ダイエット	経済，css，ビジネス，news，tmp，ネット，読み物，[，資料，mixi

表 10　(c) 意外度昇順フィルタありでの推薦上位 10 件のタグ

クエリタグ	推薦上位 10 件の共起タグ
google	checkout，evil，co-op，gears，sites，gfs，kml，appengine，starsuite，protocolbuffer
ゲーム	シューティング，クソゲー，ナムコ，ドラゴンボール改，応答，ds 攻略，ウイイレ 2008，ds ライト， 無料オンライン，ニンテンドウ
windows	sysinternals，vista，xp，7，defrag，ramdisk，azure，skydrive，cygwin，win32
firefox	add-on，add-ons，extention，拡張，機能拡張，アドオン，keyconfig，firebug，mozilla，stylish
twitter	movatwitter，favotter，jaiku，*web 文化，iran，seesmic，timelog，ヒウィッヒヒー，ツイッター，miniblog
music	daftpunk，songbird，guitar，lyrics，classic，emi，鈴木茂，細野晴臣，chaos，モーツァルト
経済	経済金融，リーマン・ブラザーズ，サブプライム，主要，keizai，subprime，景気，crisis，財政，社会病理
アニメ	ゲーマー，金田伊功，新番組，テレビアニメ，ナミ，凍結，最新刊，エグザエル，*cinema，やっぱり
アルバイト	求人，福島市，新潟市，函館市，富山市，山口市，うるま市，熊谷市，大野城市，東京 23 区
ダイエット	置き換え，寒天，痩せるダイエット，効くダイエットサプリ，脂肪を減らす薬，腹痩せ，激やせ，ウエスト， 脂肪燃焼サプリメント，黒烏龍茶

表 11　(d) 意外度昇順フィルタなしでの推薦上位 10 件のタグ

クエリタグ	推薦上位の共起タグ
google	checkout，evil，co-op，cutts]，gears，calgoo，[chrome，sites，searchwiki，co-op]
ゲーム	xbox_360，バンナム，大人の時間，アイマス p 合作動画，im@s ランク，hacchi，全ソフトカタログ， どこでもいっしょ，トロ・ステーション，obasan
windows	98，sysinternals，uac，アクティベーション，tuneup，ぴたすちお，mingw，intype，longhorn，shoutcast
firefox	検索プラグイン，tracemonkey，*拡張，add-on，userchrome，swimmie，add-ons，extention，拡張，plugin 作成
twitter	movatwitter，twit，favotter，yammer，fuuri，rubyonrail，chirrup，ふぁぼったー，jaiku，*web 文化
music	instruments，theremin，daftpunk，清志郎，matsukiayumu，rockbox，cv02，itunesstore，songbird，chuck
経済	第二次ベビーブーム，け経済ネタ，市街地，社会システム，特設：大不況，バカは罪，経済金融，＋世相， リーマン・ブラザーズ，サブプライム
アニメ	ゲーマー，かわ唯，涼宮ハルヒの消失，聖闘士星矢，とか，自主制作アニメ，2008-11-12，金田伊功，コードギアス関連，i.g
アルバイト	パートタイマー，求人，福島市，新潟市，函館市，富山市，山口市，うるま市，熊谷市，大野城市
ダイエット	☆まい☆，おぉ！いけてるじゃん♪，置き換え，女性向，寒天，体の悩み，現役モデル，バタフライフ，
	ファスティング，痩せるダイエット