社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－DBS－140 (II) (83)
2006／7／14

Folksonomyにおけるコンテンツ推薦のための メタデータ成長モデルの提案

佐々木 祥† 宮田 高道† 稲積 泰宏†† 小林 亜樹‡ 酒井 善則†

† 東京工業大学 大学院 理工学研究科 集積システム専攻
†† 神奈川大学 大学院 工学研究科 電子情報フロンティア学科
‡ 独立行政法人 メディア教育開発センター

近年急速に普及しているsocial bookmarkは，これまで利用者が個別に管理していたbookmarkをネットワーク上で他の利用者と共有するサービスである．これらのサービスでは多くの場合，利用者がbookmarkに対して「タグ」と呼ばれる自由記述のキーワードを付与することによって管理を行うFolksonomyと呼ばれる分類法を採用している．このようにして利用者が付与したタグ情報の集合は，多数の利用者によって成長を続けるメタデータであるとみなせるため，このタグ情報を利用することにより効率的なコンテンツの検索や推薦を行うことができると考えられる．しかしながら，利用者はあくまで自分自身の嗜好に基づいてタグ付けを行っているため，分類に使用したタグ名そのものは，必ずしも他の利用者とって適切なものであるとは限らない．そこで本研究では，タグの持つ本質的な情報を，タグ名ではなく，タグによるコンテンツの分類情報であると仮定し，分類間の類似度に基づいて利用者に対して望ましいコンテンツの推薦を行うシステムを提案する．

A Study of a Growing Meta-data Model for Content Recommendation on Folksonomies

Akira SASAKI† Takamichi MIYATA† Yasuhiro INAZUMI††
Aki KOBAYASHI‡ Yoshinori SAKAI†

†Department of Communications and Integrated Systems, Tokyo Institute of Technology
††Department of Electronics and Informatics Frontiers, Kanagawa University
‡National Institute of Multimedia Education

The web-based bookmark management service called social bookmark has recently come to be widely used. In the social bookmark, a user can add one or more keywords called 'tags' to their own bookmark for future use. The tag information gathering from a lot of users allows us to classify the web contents including database of social bookmark. This classification method is known as Folksonomy. Furthermore, we can qualify the tag information as the effective meta-data for search and recommendation use. However, a user hardly adds tags considering other user's convenience. Because of this behavior, a user will be not satisfied with the way to classify web contents by the name of tags. In this paper, we assume that the essential information of tags are not tag names, but classifications of web contents by tags. Based on this assumption, we have proposed the content recommendation system based on the similarities between the classifications.

1 はじめに

インターネットを介した集合知の実現形態の一つとして，ブックマーク共有，あるいはSocial bookmark（以下、SBM）とよばれるサービスが注目されている[1][2]．このサービスはWWW上で複数の利用者のブックマークを共有するサービスであり，その最大の特徴は，ブックマークにタグと呼ばれる，自由記述のキーワードを付与できることである．これに対し，ウェブディレクトリ等の従来の分類法では，サイト運営者などがトップダウン的にコンテンツを分類する必要があるため，(1) 分類のための多大なコストを運営者が負わなければならない，(2) 利用者の多様な要求に対応することが困難である，といった問題があった．

一方，利用者のタグ付与行動に基づくコンテンツの分類法は一般にFolksonomyと呼ばれており，協調的かつボトムアップ的な分類を可能にするといわれている[3]．Folksonomyは，トップダウンによる分類法に比べて，(1) 利用者に分類におけるコストを分散できる，(2) 利用者の意見に即した分類がなされる，といった利点があるとされている．

ところで，現在運営されているSBMサービスサイト[1][2]では，特定のコンテンツに対して多くの人がつけたタグを「共通タグ」と呼んでいる．このように，「ボトムアップ的な分類」は，多くのサービス利用者がコンテンツに「協調的に」タグを付与するので，少数のタグの表記のずれがあったとしても，多数決によって一意に分類が決定する，という前提でサービスが運営されている．しかしながら，実際のSBMサービス利用者は，コンテンツの分類に寄与するようなタグの付与には消極的であり，多くの場合，自分自身が使いやすいような利己的なタグ付与行動を行っているにすぎない．図1は，ある1つのコンテンツに対するタグ付与情報をタグ名ごとに集計し，多くの利用者が付与した上位のタグから降順で並べた頻度分布図である．この図に見られるように，多くの利用者が付与する上位のタグに対して，少数の利用者しか付与しない下位のタグが薄く広がっている（ロングテール現象）．そのため，多数決による分類は多くの少数意見情報を切り捨てることとなる．また，図1において付与されているタグを考察すると，「blog」「java」などの，ウェブページの大まかな内容を示すタグが上位に付けられているのに対し，下位では，「toread」（あとで読むための印付け）「web/app」（個人でカテゴリ分け）などのように，個人の管理手段を含んだタグ（以下，利己的タグと呼ぶ）が多く付けられていることがわかる．つまり，SBMの実サービスにおいては，「利用者の協調的なタグ付与」は必ずしも行われておらず，ボトムアップ的な分類がうまくいくとは限らない．

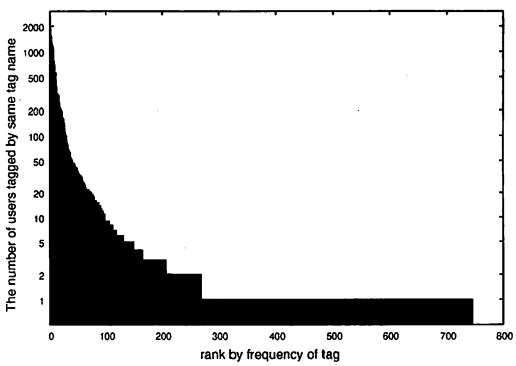

図 1: タグ付与行動の頻度分布

利己的タグは，タグ名そのものには意味がないが，特定の利用者にとっての分類手段であるため，そのタグで分類されたウェブページの集合には利用者自身にとっては何らかの共通した意味があるものと考えられる．そこで，筆者らは，多数の利用者によってコンテンツに継続的に付加され続けるメタデータについて，そのメタデータ並びにメタデータ間の関係のみを用いて，各々のメタデータをその利用に適した状態へと変化させるモデルを提唱し，これをメタデータ成長モデルと呼ぶ．本研究は，Folksonomyに分類されるサービスにおいて，多数の利用者によるタグ付け情報のみに基づいて，ある利用者のタグを表象として分類されたコンテンツ群に他のコンテンツを補完することを目的とする．本稿では，タグの表す本質的な情報であるクラスタ情報のみを用いて，分類間の類似度を計算し，これに基づいた分類の統合とコンテンツの補完を行うことで，目標とするタグに対応するコンテンツを推薦する手法，並びにアルゴリズムを提案する．

本手法では，協調フィルタリング[4]を技術的要素として用いる．協調フィルタリングは，商用サイトなどで用いられる商品推薦システムであり，大別すると，利用者ベース[5]のものとアイテムベース[6]のものがある．利用者ベースの協調フィルタリングでは，利用者がアイテムに対して明示的，あるいは暗示的に評価を行い，類似した評価を持つ利用者グループを探し出し，類似する利用者グループが高評

価を与えたアイテムを推薦する手法である．これに対し，アイテムベースの協調フィルタリングでは，アイテム同士の評価の類似性を基準にアイテムを推薦する手法である．本手法では，利用者ベースの協調フィルタリングを拡張し，ある利用者のタグを表象として分類したコンテンツ群をベースとした，いわばカテゴリベースの協調フィルタリングを提案する．

文献 [7] ではブックマークを利用したウェブページ推薦システムが提案されている．この手法は，個人のブックマークのカテゴリ分けに基づいて類似するウェブページを推薦するシステムである．しかしながら，今回対象とする SBM のタグ情報は，図 2 に示すように複数のタグによって重複を許した分類となっているため，問題はより複雑となっている．文献 [8] では，SBM を利用したリコメンデーションシステムが提案されている．この研究では，意味が類似していると判断したタグをクラスタ化することで利用者ごとの分類表現のゆらぎの影響を軽減している．しかしながら，この研究は，タグ名による推薦システムであり，利用者毎のタグの意味の違いには着目していない．

2 研究概要

2.1 SBM の定式化

まず，利用者とタグの関係について定式化を行う．SBM サービスの利用者に $u_1, u_2, \cdots$ のように番号を与える．また，u_i が SBM において利用しているタグの集合を $\langle u_i \rangle$ と記述し，SBM で利用されているタグ t_j を次のように通し番号を与える．

$$\langle u_1 \rangle = \{t_1, t_2\}\ ,\ \langle u_2 \rangle = \{t_3, t_4, t_5\}\ ,\ \langle u_3 \rangle = \{t_6, t_7\} \cdots \quad (1)$$

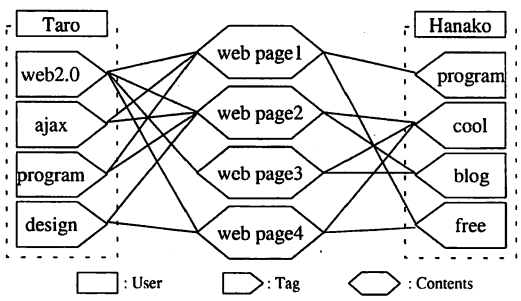

図 2: SBM モデル

つまり，利用者 u_i が SBM において使用しているタグを，以下の式が成り立つように番号を与える．

$$\langle u_i \rangle = \{t_{o(i)}, t_{o(i)+1}, \cdots, t_{o(i+1)-1}\} \quad (2)$$

ただし，$o(i)$ は，タグの通し番号であり，u_i の最初のタグの番号である．すなわち，たとえある 2 人の利用者が同一のタグ名を利用していた場合であっても，本手法では異なるタグとして区別されることになる．以下，ある特定の利用者 u_i によるタグ t_j を強調するときは，これを「ユーザタグ」と呼び，$u_i : t_j$ と記述する．

次に，ユーザタグとコンテンツの関係について定式化を行う．SBM に登録された全コンテンツ集合を A(All data) とし，利用者 u_i が自身のブックマークとして登録したコンテンツ集合を $B(u_i)$ とする (Bookmark)．また，集合 $B(u_i)$ のうち，利用者 u_i がユーザタグ $u_i : t_j$ を付与したコンテンツ集合を $T(u_i : t_j)$ とする (Tagged)．ただし，ユーザタグはひとつのコンテンツに対して複数付与することが出来るので，$\{T(u_i, t_1), T(u_i, t_2), \cdots\}$ は重複を許した集合となる．

ここで，あるユーザタグ $u_i : t_j$ に着目してコンテンツを分類したとき，SBM に登録された全コンテンツは図 3 のように分けられ，$A \cap \overline{B(u_i)}, B(u_i) \cap \overline{T(u_i : t_j)}, T(u_i : t_j)$ のいずれかに属することとなる．$A \cap \overline{B(u_i)}$ は，u_i がブックマークしておらず，かつ他の SBM 利用者がブックマークをしているコンテンツ，$B(u_i) \cap \overline{T(u_i : t_j)}$ は，u_i がブックマークしたコンテンツの内，タグ t_j を付与していないコンテンツ，$T(u_i : t_j)$ は，u_i がブックマークし，かつタグ t_j を付与したコンテンツを意味する．SBM 内のコンテンツはすべて $c_1, c_2, \cdots$ のように識別子が与えられているものとし，利用者 u_i のブックマークと，同利用者のユーザタグ $u_i : t_j$ による分類の結果，コンテンツ c_k がいずれに属するか，というコンテンツの属性を $E(u_i : t_j, c_k)$ で表すこととし，以下のように定め

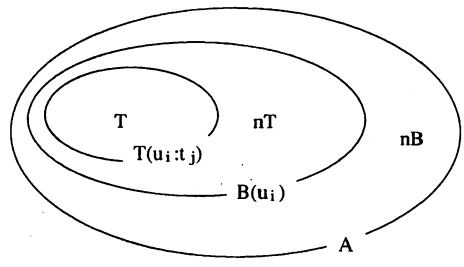

図 3: SBM におけるコンテンツ集合の関係

る (図 3).

$$E(u_i:t_j,c_k)=\begin{cases} T & :c_k\in T(u_i:t_j) \\ nT & :c_k\in B(u_i)\cap\overline{T(u_i:t_j)} \\ nB & :c_k\in A\cap\overline{B(u_i)} \end{cases} \quad (3)$$

また，ユーザタグ $u_i:t_j$ において，コンテンツ c_k の属性を k 番目の成分とするベクトルを考えることが出来る．これをタグベクトル $\vec{t_j}$ と呼ぶこととし，次式で定める．

$$\vec{t_j}=(E(u_i:t_j,c_1),E(u_i:t_j,c_2),\cdots) \quad (4)$$

2.2 提案システムの概要

今回想定するアプリケーションは，利用者 u_i が SBM において用いているタグ t_j を入力として，そのタグ t_j に分類されるべきコンテンツを推薦するものである．まず，コンテンツの推薦は，次のフローで行われる（図 4）．

1. 利用者 u_i がタグ名 t_j をシステムに入力
2. 利用者の SBM の履歴からタグベクトル $\vec{t_j}$ を生成
3. 提案するアルゴリズム (3 章) によって，タグベクトル $\vec{t_j}$ を補完
4. アルゴリズムによって補完されたコンテンツを推薦

次に，推薦基準について説明する．図 5 は，利用者 u_1 がユーザタグ $u_1:t_1$ に類似するコンテンツを検索した結果，他の利用者 u_2 のユーザタグ $u_2:t_2$ からコンテンツ c_7 が推薦される概念を示している．

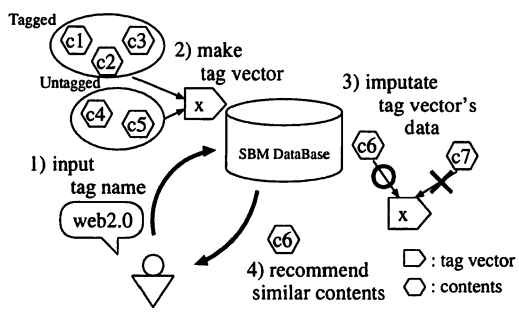

図 4: 提案推薦システムの概略

$u_1:t_1, u_2:t_2$ におけるタグベクトルは次の式で示される．

$$\vec{t_1}=(T,nT,T,T,T,nT,nB,nT,nB) \quad (5)$$

$$\vec{t_2}=(nB,nB,T,T,T,nT,T,nT,nB) \quad (6)$$

ここで，2 つのタグベクトルを比較し，タグベクトルの類似性を計算する．どちらかのタグベクトルで nB である成分は，比較ができないため類似性の判定基準から除外し，それ以外の成分において属性が多く一致するとき，2 つのタグベクトルは類似性があると判定する．よって，$\vec{t_1}$ と $\vec{t_2}$ の類似性は高い．そのため，$u_1:t_1$ に含まれておらず，$u_2:t_2$ に含まれているコンテンツ c_7 は，u_i が閲覧したときに，タグ t_1 を付与してブックマークをする可能性が高いと考えられるので，コンテンツ c_7 を推薦する．

3 提案法

タグベクトルの補完のアルゴリズムについて説明する．このアルゴリズムは，以下の過程で構成される（図 6）．

1. タグベクトル間の類似度計算
2. タグベクトルの補完度計算
3. タグベクトルの補完

3.1 タグベクトル間の類似度計算

まず，タグベクトル間の類似性の度合いを表す類似度関数 sim を導入する．タグベクトル間の類似度は以下の式によって算出する．

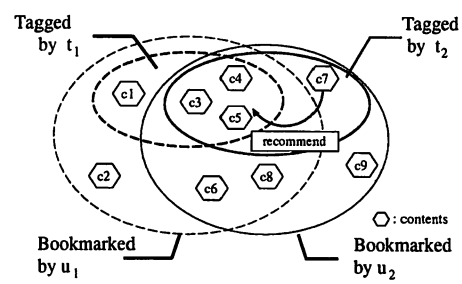

図 5: コンテンツの推薦概念

表 1: タグベクトルのサンプルデータ

user:tag	tag name	c_1	c_2	c_3	c_4	c_5	c_6
$u_1 : t_1$	hobby	0	0	1	1	nB	0
$u_1 : t_2$	web2.0	0	1	1	0	nB	0
$u_1 : t_3$	toread	1	1	0	0	nB	1
$u_2 : t_4$	fun	0	0	1	1	1	nB
$u_2 : t_5$	news	1	1	0	0	0	nB
$u_3 : t_6$	blog	1	0	0	1	0	nB
$u_3 : t_7$	toread	0	1	0	1	1	nB

表 2: サンプルデータにおける補完計算結果

user:tag	tag name	c_1	c_2	c_3	c_4	c_5	c_6
$u_1 : t_1$	hobby	0	0	1	1	0.75	0
$u_1 : t_2$	web2.0	0	1	1	0	0.67	0
$u_1 : t_3$	toread	1	1	0	0	0.25	1
$u_2 : t_4$	fun	0	0	1	1	1	0.00
$u_2 : t_5$	news	1	1	0	0	0	0.67
$u_3 : t_6$	blog	1	0	0	1	0	0.50
$u_3 : t_7$	toread	0	1	0	1	1	0.33

$$\mathrm{sim}(\overrightarrow{t_x}, \overrightarrow{t_y}) = \frac{\overrightarrow{t_x} \cdot \overrightarrow{t_y}}{|\overrightarrow{t_x}||\overrightarrow{t_y}|} \tag{7}$$

ただし，前述したように，この計算では，比較を行う 2 つのタグベクトル t_x, t_y において，N を含む成分を除外する．また，$T = 1, nT = 0$ として計算する．

3.2 タグベクトルの補完度計算

次に，タグベクトルの類似度を用いて，タグベクトルの補完度計算をする．補完度とは，タグベクトルのうち nB である成分を補完する際の基準となる値である．ここでは，補完するタグベクトルを t_x とし，補完対象となる成分はの k 番目であるとし，求める補完度を P_{xk} とする．以下の計算式によって，コンテンツ c_k に対するタグベクトル t_x との補完値 P_{xk} を算出する．

$$P_{xk} = \frac{\sum^y \mathrm{sim}(\overrightarrow{t_x}, \overrightarrow{t_y}) \cdot H_{yk}}{\sum^y \mathrm{sim}(\overrightarrow{t_x}, \overrightarrow{t_y})} \tag{8}$$

ここで，比較対象となるタグベクトル群 $\{\overrightarrow{t_y}\}$ は，SBM 内のタグベクトルの集合から k 番目 N でない成分から選択される．

3.3 タグベクトルの補完

3.2 において求めた補完度は，0 から 1 までの値をとる．ここで，実際に該当コンテンツを推薦するかどうかを閾値を設けることで決定する．すなわち，閾値を th とし，以下の式が成り立つとき，コンテンツを推薦する．

$$P_{xk} \geq th \tag{9}$$

4 適用事例

4.1 サンプルによる計算例

表 1 はタグベクトルのサンプルである．例えば，利用者 u_1 によるタグ「hobby」は，推薦システムにおいてはタグベクトル t_1 として以下のように管理されている．

$$\overrightarrow{t_1} = (0, 0, 1, 1, nB, 0) \tag{10}$$

次に，t_1 と t_6 との類似度 $\mathrm{sim}(\overrightarrow{t_1}, \overrightarrow{t_6})$ は以下のように計算される．

$$\mathrm{sim}(\overrightarrow{t_1}, \overrightarrow{t_6}) = \frac{(0,0,1,1) \cdot (0,1,0,1)}{|(0,0,1,1)||(0,1,0,1)|} = 0.5 \tag{11}$$

この類似度を元に，タグベクトルとコンテンツの関係度を計算する．以下は，t_1 と c_5 の関係度 P_{15} を，(コンテンツ c_5 に対応する) 5 番目の成分が nB でないタグベクトル t_4, t_5, t_6, t_7 から算出したものである．

$$P_{15} = \frac{1 \cdot 1 + 0 \cdot 0 + 0.5 \cdot 0 + 0.5 \cdot 1}{1 + 0 + 0.5 + 0.5} = 0.75 \tag{12}$$

以上の計算を行った結果を表 2 に示す．

この結果から，閾値を決めることで，推薦を行うかどうかを決定する．例えば，閾値を 0.6 以上と決めることで，t_1, t_2 には c_5 が，t_5 には c_6 が推薦される．

4.2 考察

表 2 は，表 1 のサンプルに対し，提案法を適用した結果である．利用者 u_1 がタグ「hobby」で分類したコンテンツ集合は利用者 u_2 がタグ「fun」で分類したコンテンツ集合に近いため，利用者 u_1 にコンテンツ c_5 を推薦している．このように，提案法では，タグ名の関連性を見ることなく「hobby」と「fun」との関連性をコンテンツ集合の共起性から読み取り，コンテンツの推薦を行うことが可能となる．

本手法の特徴的な結果として，u_2 による「news」と u_1 による「toread」のように，言葉の意味として

はおおよそ関連性がない場合についても，コンテンツ集合の共起性からコンテンツの推薦が行えることがあげられる．逆に，同じタグ名である「toread」に関して，u_1とu_3とでは重要とする事柄が異なるため，異なるコンテンツ集合となり，「toread」同士ではリコメンドされない．これは，文献 [8] にようなタグ名による推薦機構では実現できないことであり，本手法のようなタグの表す本質的な情報であるクラスタ情報に基づいた推薦機構だから達成できることである．

5 おわりに

本稿では，タグ名そのものには注目せず，ある利用者によって同じタグを付けられたコンテンツ集合にこそ意味があるという仮定に基づいて，利己的なタグ付け行動を行う利用者に対しても類似コンテンツを推薦できる手法を提案した．この成果によって，「自由記述によるブックマークの管理」が可能なSBMの利点を確保しつつ，「ボトムアップによる分類」というFolksonomyのコンセプトが達成できると考える．

今回は，利用者同士のSBM活動履歴を用いたウェブページ推薦機構として提案したが，システム側が予めタグベクトルをメタデータ化して準備しておくことも可能である．類似度の高い複数のタグベクトルを用い，コンテンツの出現確率をもとにタグベクトルを生成することにより，推薦に必要とされる計算量の削減が可能になると思われる．

今後の検討課題としては，「blogでかつnews」，「blogまたはnews」のような複数のタグにまで比較対象を広げたand，or指定への拡張が挙げられる．SBMの実サービスサイトにおけるタグ付け傾向を見ると，「web」「app」と分けてタグ付けを行う利用者がいる一方で，「web/app」等の個人での階層化を行う利用者も見られる．また，厳密に多くのタグを用いて分類する利用者もいれば，大雑把に少ないタグを用いて分類する利用者もいる．そのため，より精度の高い推薦システムを構築する上で，and，orまで含めて比較する必要ではないかと考えられる．and，orの拡張においては，比較対象が莫大に増え，計算量が多くなることなどが予想されるため，今後は，前述したタグベクトルのメタデータ化とともに検討していく必要があると考えられる．

参考文献

[1] del.icio.us，http://del.icio.us/

[2] はてなブックマーク，http://b.hatena.ne.jp/

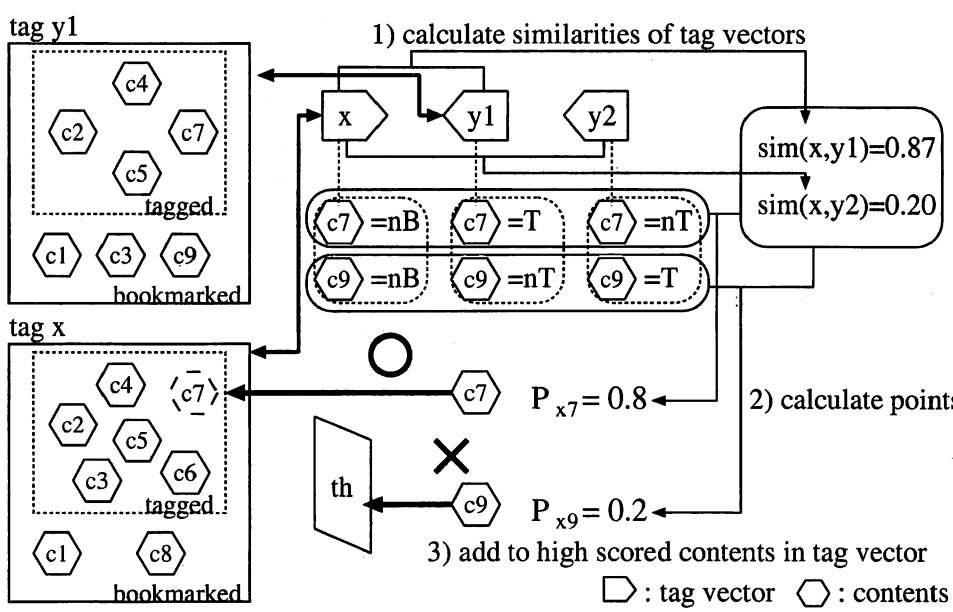

図 6: 提案アプリケーションの概念

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

[4] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, "Using collaborative filtering to weave an information tapestry," Communications of the ACM, Vol. 35 No. 12, pp. 61–70, Dec. 1992.

[5] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," Proceedings of the 1994 Computer Supported Cooperative Work Conference, pp. 175–186, 1994.

[6] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. "Item-based collaborative filtering recommendation algorithms," Proceedings of the 10th International World Wide Web Conference (WWW10), May 2001.

[7] J. Rucker, M.J. Polanco, "Siteseer: Personalized navigation for the web," Communications of the ACM, Vol. 40, No. 3, pp.73–75, 1997.

[8] S. Niwa, T. Doi, S. Honiden, "Web Page Recommender System based on Folksonomy Mining," Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06), 2006.