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

2007－DBS－143（68）
2007／7／3

Folksonomy のタグを用いた自動分類体系構築へ向けて

江田 毅晴† 吉川 正俊†† 山室 雅司†

† NTTサイバースペース研究所 〒239-0847 神奈川県横須賀市光の丘 1-1
†† 京都大学情報学研究科 〒606-8501 京都市左京区吉田本町
E-mail: †{eda.takeharu,yamamuro.masashi}@lab.ntt.co.jp, ††yoshikawa@i.kyoto-u.ac.jp

あらまし 本研究では, Folksonomy データの分析に基づく, 新しい分類体系構築方法について提案する. 現在のソーシャルブックマークサービスでは, 大量のブックマークエントリの中から有用なリソースを探すには, キーワードやタグを指定するか, 興味の近いブックマーカを努力して探し出す必要がある. 本研究では, タグの共起関係に基づく意味的な繋がりを利用して, タグの集合を分類構造として体系だてる. これにより, 利用者にはタグの関連を認識した直観的な探索を通して, 有用なリソースを探しだすことが可能となる.

キーワード ソーシャルタギング, ソーシャルブックマーク, フォークソノミー, CSCW

Towards Automatic Web Page Classification based on Tags in Folksonomies

Takeharu EDA†, Masatoshi YOSHIKAWA††, and Masashi YAMAMURO†

† NTT Cyber Space Laboratories 1-1 Hikarinooka, Yokosuka-Shi, Kanagawa, 239-0847 Japan
†† Department of Social Informatics, Kyoto University Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan
E-mail: †{eda.takeharu,yamamuro.masashi}@lab.ntt.co.jp, ††yoshikawa@i.kyoto-u.ac.jp

Abstract In this research, we propose a new classification system based on the analysis folksonomy data. In order to find valuable resources from current social bookmark services, users need to specify search terms or tags, or to discover people with the similar interests. Our proposal utilizes semantic relationships extracted from the cooccurrences folksonomy data and organizes a new classification system from tag sets. Users can find useful resources by navigating tag relationships intuitively.

Key words Social Tagging, Social Bookmark, Folksonomy, CSCW

1. はじめに

昨今, ブログやウィキといったユーザ参加型のウェブサービスが普及している. これらのサービスの大きな特徴として, 書き込みの頻度が非常に高く, ページ間のリンクは半自動的に生成されることが多い. その結果, 従来のリンク解析に基づく Web コンテンツのランキングが難しくなりつつあるという現状がある.

一方, folksonomy という仕組みに基づく ソーシャルブックマークサービス ([2], [8], など) が注目を集めている. ソーシャルブックマークサービスでは, 参加ユーザのブックマークをリモートサーバに置いて共有する. ブックマークした人数を利用してウェブサイトのランキングを提供したり, ブックマークへのコメントを通しての緩いコミュニティ形成を実現している.

Folksonomy の 1 つの特徴として, **タグ** (tag) がある. タグとは, 各ユーザがリソースをブックマークする際の切り口を与える方法であり, 複数のキーワードの集合をブックマークに付与することが出来る. 付与されたタグは, 後日自分のブックマークを探したいときのフィルタ条件として利用することができる. タグはユーザが任意に選べ, 必ず付与しなければいけないというような強制も無いため, タグを付与することへの敷居は非常に低い. この結果, 研究はされていたもののなかなか広まらなかった, ウェブリソースへのアノテーションを爆発的に普及させることになった.

タグを分類方法として捉えた場合, 分類の専門家を雇う人件費や時間をかけてリンク解析をする必要もないため, ブログエントリ等の鮮度の高い情報の分類には非常に適している. こうした意味で, ソーシャルブックマークサービスは, 次世代のウェブリソースのランキングおよび分類方法の有力候補の一つと考えられる.

しかしながら, folksonomy にもいくつか問題点がある. 同じ意味を表していても人によってタグが異なるという類義タグの問題や, 1 つのタグで複数の意味を持つ多義タグの問題である.

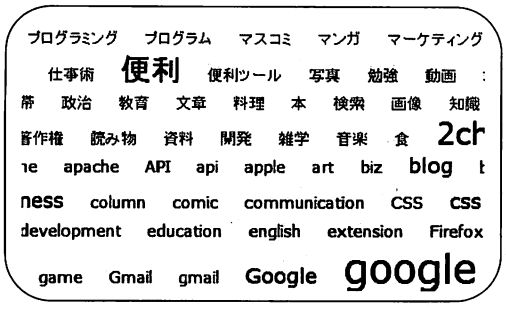

図 1 五十音順タグクラウドの一部.

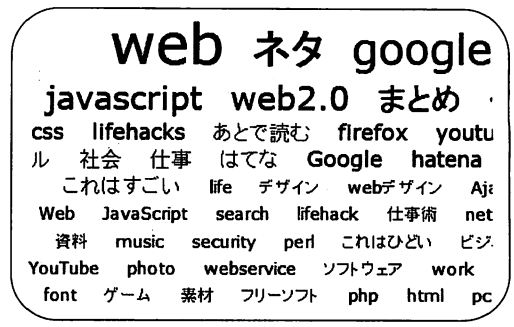

図 2 利用頻度順タグクラウドの一部.

これらの問題に対し, Xian 等は, folksonomy データを統計的に解析することによって, 部分的な解決手法を提案している [7]. 本研究では, Xian 等の索引付け手法を利用し, folksonomy のタグを整理し, 直観的に理解可能で, ウェブリソース探索のためのもう一つの手段として使える分類体系の構築を目指す. Xian 等の手法を「部分的」と言う理由は, 彼等の論文では, 類義タグ多義タグの特徴ベクトルの傾向を示し, 確率的な検索システムを提案したに過ぎず, 実際の検索時の問題の解決方法にまでは触れていないからである. Xian 等の提案した, PLSI に基づく folksonomy の索引付け手法は類義, 多義タグ問題の完璧な解決手段ではない. 本稿では, 実データを利用してどれ位の精度まで達成できるのかを確認した上で, その特性を理解した上で分類体系として利用する具体的方法を提案する.

本稿の構成は下記のようになる. まず既存手法であるウェブディレクトリとタグクラウドについて説明する. 続いて, PLSI を用いて folksonomy データの索引付けについて説明し, 三つの提案手法について説明する. 最後にまとめと今後の課題について述べる.

2. ウェブディレクトリと単純なタグクラウド

ウェブページの分類方法として古くから利用されているのがウェブディレクトリである. ウェブディレクトリでは, 分類の専門家を雇ってウェブページ群をファイルシステムのような階層的な分類構造のどこかのノードに位置づける. 専門家が熟考して分類を決定するため, 分類の精度が良いことが期待できる. また, ウェブページを探す際にも, ディレクトリを辿ることにより抽象度を下げるという, 直観的に理解しやすい探索インタフェースを提供することが出来る. ウェブディレクトリの課題としては, 分類出来るページ数に限界があること, 最新のブログエントリやニュース記事, 掲示板のスレッドのような鮮度の高い情報を分類することが困難であるといったことが挙げられる.

一方で, ソーシャルブックマークサービスにて提供されているタグ一覧表示方法として, タグクラウドが存在する. タグクラウドとは, ソーシャルブックマークサービス内で利用されているタグのうち利用頻度が高いものを抽出し一覧表示したものであり, 利用頻度が大きいほどフォントのサイズを大きく表示している. タグの並びの順序として, 五十音順 (図 1) と頻度順 (すなわちフォントの大きさ順)(図 2) が利用されている.

しかしながら, 既存のタグクラウドではタグが並んでいるだけであるため, 実際にこれを利用して何らかの情報を探そうとすると, 似たようなタグを何度も探して試す必要があることに気づく. これは, 一覧表を示しているのにも関わらずタグの並びに意味的な繋がりがないためである.

3. 提案手法

上記の検討を踏まえ, 我々は PLSI を用いてタグを索引付け, それぞれのタグの特徴ベクトルを利用して, folksonomy システムから自動的に生成される新しい分類体系を提案する. 検討過程を説明しながら, 次の 3 段階で分類体系をブラッシュアップしていく.

(1) グループタグクラウド
(2) 自動カテゴリ
(3) 自動カテゴリの洗練

研究の前提

本研究では, タグの文字列としての特徴や辞書, 人手による判断は出来るだけ行わずに, ソーシャルブックマークのデータを単純な三つ組からなる共起 (Co-occurrence) データとしてのみ利用する手法を構築することを目標とした. 実際のサービス提供時には, 辞書や人手での判断を適宜追加することにより, 分類体系としての完成度をさらに上げることができる.

3.1 PLSI を用いた folksonomy データの索引付け

PLSI(Probabilistic Latent Semantic Indexing) を用いると, 複数のアイテム集合の間での共起関係データ (Co-occurence Data) から, 与えられた次元数の確率値ベクトル空間へそれぞれの集合のアイテムを射影することが出来る.

図 3 にモデルを示す. ユーザ集合, リソース集合, タグ集合をそれぞれ, U, R, T とする. 今, 射影したいベクトル空間の次元数を D, ある次元 $\alpha(1 \leq \alpha \leq D)$ からユーザ $u_i \in U$, リソース $r_j \in R$, タグ $t_k \in T$ の生起確率を, それぞれ $p(u_i|\alpha), p(r_j|\alpha), p(t_k|\alpha)$ とする. また次元 α の生起確率を $p(\alpha)$ とすると, u_i, r_j, t_k の間には, 潜在的意味ベクトル空間を介在して以下の関係式が成り立つ(注1).

$$p(u_i, r_j, t_k) = \sum_{\alpha=1}^{D} p(\alpha)p(u_i|\alpha)p(r_j|\alpha)p(t_k|\alpha)$$

(注1) : 但し, $p(u_i, r_j, t_k)$ は, 三つ組の共起確率.

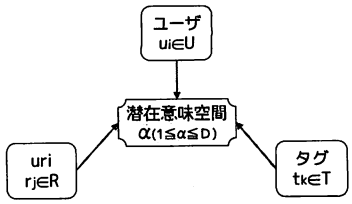

図3 モデル.

左辺はデータセットから観測出来る値であり，右辺の確率値は適当な初期値を与えることによって，EMアルゴリズムによって計算可能である．本手法では，温度パラメータを与えて，平滑化を行ったEMアルゴリズムを用いて計算している[9]．

次にベイズ則を用いて，各潜在意味次元からユーザ，リソース，タグへの条件付き確率を計算することが出来る．α 次元からの条件付確率を，それぞれのアイテムの α 次元の値とすることによって，すべてのアイテムを D 次元ベクトルとして表現出来る．以降，PLSIにて索引付けされたベクトルをPLSIベクトルと呼ぶことにする．

3.1.1 PLSIベクトルの観察

図4にPLSIベクトルの例を示す．横軸が次元であり，縦軸がその次元の値である．図4にあるPLSIベクトルは，主に「ソーシャルブックマーク」に関するタグを表示しているが，全てのベクトルが，第9次元に強く特徴を持っていることが分かる．

アイテム頻度

アイテム頻度とは，三つ組データから，どれかの属性を指定して選択した三つ組の濃度のことを指す．

$$\text{アイテム頻度}_{u='\text{松木}'}^{t='\text{サッカー}'} = \{\ '\text{松木}'\text{さんが}'\text{サッカー}'\text{タグを利用した回数}\ \}$$

Folksonomyデータの場合，指定属性が三つで，三つともそれぞれの集合中に値が存在した場合，アイテム頻度は1になる．指定属性が一つでタグの場合，特にタグアイテム頻度と呼ぶ．

PLSIベクトルのエントロピー

タグ t_k のエントロピー値とは，以下の式で計算される値である(注2)．

$$H(t_k) = -\sum_{\alpha=i}^{D} p(\alpha|t_k) \cdot log(p(\alpha|t_k))$$

エントロピーの値は，PLSIベクトルが1つのピーク(最大値=1)を持つ場合には0をとり，一様分布に近いほど高くなり，PLSIベクトルの曖昧性を表していると考えることが出来る．図5には，全PLSIベクトルの中から，エントロピー値が高いPLSIベクトルを示している．図6は，逆にエントロピー値が0であるPLSIベクトルを表示している．

PLSIベクトル間の距離

なお，本研究では距離尺度として確率値ベクトル間での距離として有効であるとされるJSダイバージェンスを用いている．JSダイバージェンスは，下記の式で計算される(注3)．

$$D_{js}(t_k, t_l) = \frac{1}{2}[D(t_k||avg(t_k, t_l)) + D(t_l||avg(t_k, t_l))]$$

距離尺度の選択については，特に異なるアイテム間での距離を測る際に，議論の余地がある[4]．

表1 実験データ.

三つ組数	674619
ユーザ数	3065 人
リソース数	3353 個
タグ数	1201 個

表2 PLSIのパラメータ.

潜在次元数	80 次元
EM 繰返し回数	80 回
最低アイテム頻度	50 回

3.2 予備評価

我々は，PLSIベクトルの距離の近さが近いほど，タグが同義である可能性は高いと予想した．また，タグ間のis-a関係は，全体の中でエントロピーが高いタグを上位にとり，下位に上位タグに距離が近くエントロピーが低いタグを設定することで構成できると予想した．

そこで，これらの条件によりどれくらいの精度で判定が出来るかどうか確認するために予備実験を行った．

3.2.1 実験方法

同義関係

アイテム頻度の高いタグを適当に選択し，それぞれのタグから距離の近いタグを複数取得する．距離の近いタグは，距離に対する閾値を設定して取得する．閾値内に含まれるタグが無い場合は同義タグが存在しないとして評価対象には含まない．距離の近いタグが同義タグであるかどうかを判断し，正答率を計算する．

is-a 関係

アイテム頻度が高く，かつそれらのうちPLSIベクトルのエントロピーの高いタグを複数取得する．それぞれのタグを上位タグとみなして，距離の近い下位タグを取得する．作成された上位下位関係が実際に，is-aとみなせるかどうかを判断し，正答率を計算する．

実験データセットとPLSIのパラメータ

2006年10月頃におけるはてなブックマーク[8]の人気エントリ集合を利用した．はてなブックマークのタグ一覧(`http://b.hatena.ne.jp/t/`)にあるタグを取得し，そこからたどれるurl集合から，三つ組データを取得した．パーズし，最低アイテム頻度でフィルタした後の実験データは表1のようになる．最低アイテム頻度を設定するのは，ノイズを除去するためである．最低アイテム頻度が低い場合は，ノイズの影響を強く受けるため全体的に精度が落ちることが分かっている．PLSIのパラメータは，表2のように設定した(注4)．

3.2.2 結果

同義関係

タグのPLSIベクトルから指定した距離の範囲内に含まれるタグのうちの同義タグの割合を正解率とした

(注2)：ユーザ，リソースにも同様に定義される．

(注3)：$D(q||r)$ はKLダイバージェンス．

(注4)：これらのパラメータは関連研究[7]を参考に独自にチューニングしたものである．

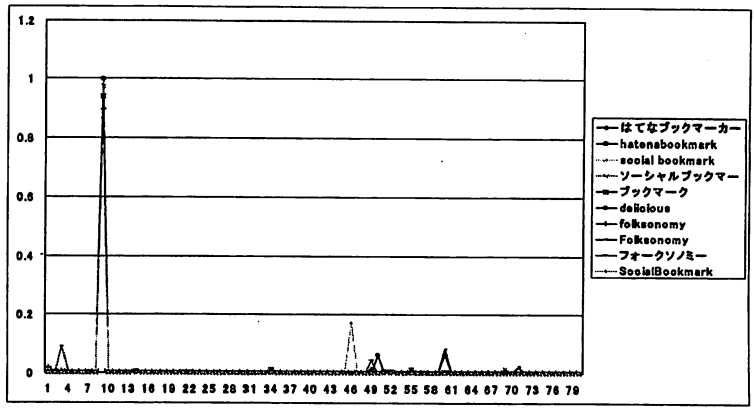

図 4 PLSI ベクトルの例.

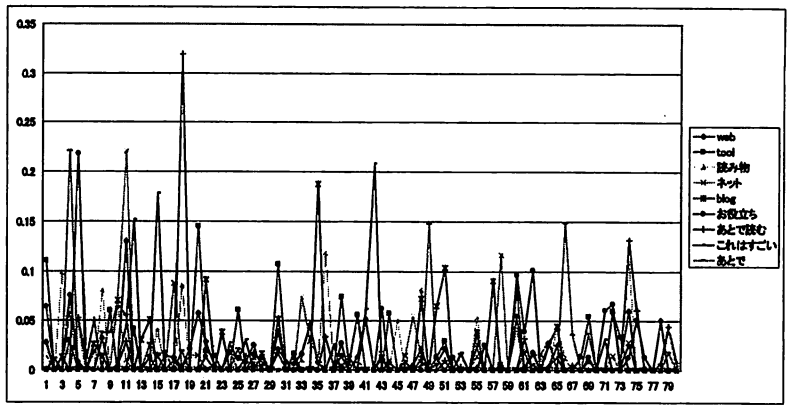

図 5 エントロピー値の高い PLSI ベクトルの例.

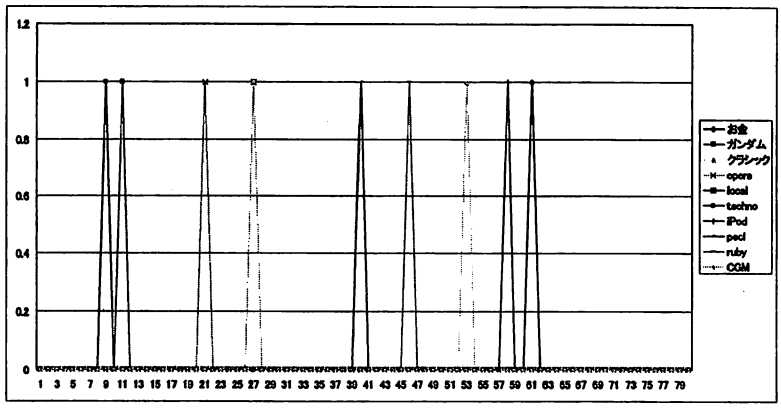

図 6 エントロピー値が 0 の PLSI ベクトルの例.

場合の平均値を図 7 に示す. 距離の閾値としては, [0.001, 0.005, 0.01, , 0.03, 0.05, 0.07, 0.1] について示しており, この結果は正解集合の再現率ではないが, 近距離に含まれるタグのうち何割かには同義タグが含まれることが多いことを示している.

is-a 関係

エントロピーが高いタグの近くにはそれ程多くのタグは集まらないため, 距離の閾値を変更し, [0.1, 0.2, 0.3, 0.4, 0.5] にした. 図 8 に結果を示す. 同義タグの平均正解率は, 距離が小さいところで頭打ちになっているのに対して, is-a タグの平均正解率は上昇している. これは, 上位タグ候補として, エントロピーの高いタグを選択しているためであり, ある程度までは上昇傾向

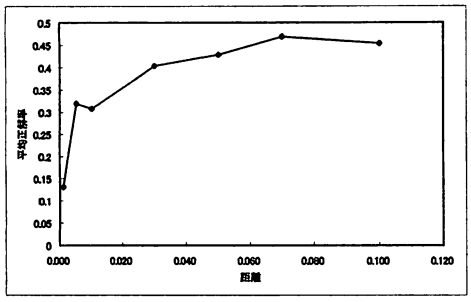

図 7　距離範囲内に含まれるタグの平均正解率 (同義タグ率).

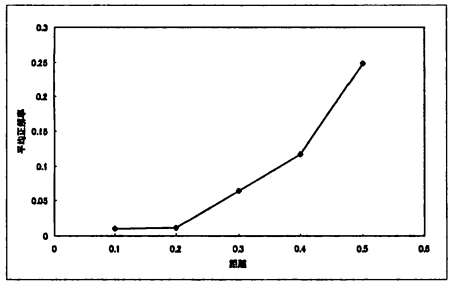

図 8　距離範囲内に含まれるタグの平均正解率 (is-a タグ率).

思想, gender, ジェンダー, 批評, society, 歴史, 考察, 差別, 中国, 哲学, china , korea, 韓国, 時事, 宗教, カルト, 外交, 選挙, 政治, religion, 北朝鮮, politics, 社会, いじめ, world, 国際, 心理, history, 心理学, 安倍晋三, 議論, 犯罪, アメリカ, 事件, ゲーム理論, 行政

図 9 「興味深い」に近いタグ一覧.

は継続すると予想される.

考察

予備実験により, PLSI による索引付けのみで, 同義タグのおよび is-a 関係を判定できる可能性はそれ程高くないと言える.

しかしながら, 予備評価の段階で, 通常のウェブディレクトリ等の分類体系やタグクラウドでは決して見ることの出来ないと思われるタグ間の関連を抽出できることもあった. 例として, 「興味深い」というタグに近くエントロピーが低かったタグ一覧を図 9 に示す. このタグ一覧は上位から機械的に取得したものである.

本研究の目標は, folksonomy データから直観的な分類体系を構築することにあるが, 硬い情報を取得することだけを目標とするのは, 新しい分類体系を提案するという可能性を狭めてしまう恐れがある. そこで我々は, ある程度の硬さを与えつつも多少の誤差を残すことは, 新しい技術を構築する上で, 必ずしも避けるべきことではないと判断し, 誤差を残しつつもいくつかのタグの分類技術を構築することとした.

3.3　グループタグクラウド

グループタグクラウドは, タグの PLSI ベクトルを K 平均法にてあらかじめクラスタリングしておき, 同一クラスタに含まれるタグは同一のグループとしてタグクラウドを構成したものである. 作成手順は下記のようになる.

（1） タグ集合全体を利用してクラスタリングを行い, 各タグにクラスタ ID をあらかじめ付与しておく.

（2） タグ集合のうち, タグアイテム頻度が上位 k 件に入るタグを取得する.

（3） 取得された上位タグをクラスタ ID に沿ってグループ分けし, 各グループ内での最大頻度タグのタグ名で五十音順にグループを整列する.

k はシステムの要求に応じて適宜決定することが可能である. 図 10 に例を示す. 意味の近いタグは同じグループにまとめられることにより, 全体の並びを直観的に理解することが可能となっている.

図 10　グループクラウドの一部.

3.4　自動カテゴリ

自動カテゴリは, エントロピー値を用いてタグ間の is-a 関係を抽出し, 分類構造として利用することを目指して設計した. PLSI ベクトルのエントロピー値が高いほど曖昧性が高いという観察に基づき, エントロピー値の高いタグを上位概念, そのタグに近いタグを下位概念として表示する. 作成手順は下記のようになる.

（1） タグ集合のうち, タグアイテム頻度が上位 k 件に入るタグを取得する.

（2） 取得したタグ集合をエントロピー値に基づいてソートする.

（3） エントロピー値が高い順に, それぞれのタグから近距離タグを展開する.

図 11 に, 上記手順にて作成した自動カテゴリの一部分を示す. 自動カテゴリでは, エントロピーの高いタグを上位に持ってくることにより, タグ一覧に対して, タグ間の親子関係を辿ることにより抽象度を下げるという直観的な理解を提供することができる.

3.5　自動カテゴリの洗練

図 11 からも分かるように, 単純な自動カテゴリには, 下記の問題がある.

（1） 下位タグの中に上位タグの同義タグが混入してくる

（2） 上位タグ全体の中に同義タグが登場することがある

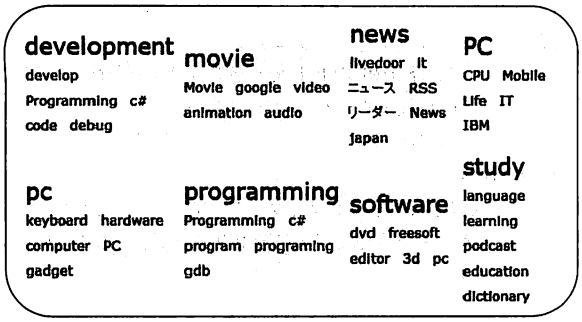

図 11 自動カテゴリの一部.

(3) 「あとで読む」や,「これはすごい」といった分類基準としてはあまりふさわしくないタグが上位タグに多い.

これらの課題を, グループタグクラウドと自動カテゴリのアイディアを混ぜ合わせることで, 自動カテゴリを改良する. グループタグクラウドの結果を観察すると, 同義タグは頻度の高いタグの間では, 同一グループに属することが多い. この知見を利用して手順を改良したものが下記になる.

(1) タグ集合全体を利用してクラスタリングを行い, 各タグにクラスタ ID をあらかじめ付与しておく.

(2) タグ集合のうち, タグアイテム頻度が上位 k 件に入るタグを取得する.

(3) 取得したタグ集合をエントロピー値に基づいてソートする.

(4) エントロピー値が高い順に, それぞれのタグから近距離タグを展開するが, 上位タグ候補は全体の中で 1 グループに 1 つのみとなるようにする.

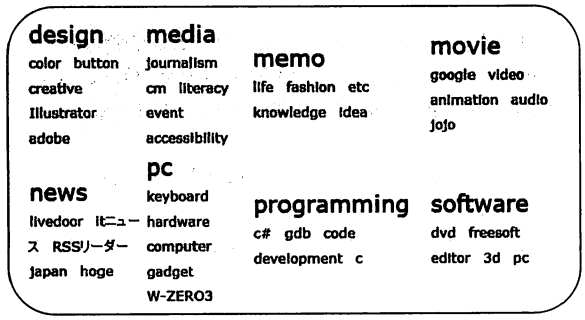

図 12 洗練された自動カテゴリ.

4. 関連研究

昨今, この分野は非常に注目を集めており, 自動タグを利用して階層を構築している [1] や, del.icio.us の非常に高頻度のタグを利用して階層構造を構築している [3] 以外にも多数の関連研究がある. オントロジーとタグの関連については [5] が分かり易く, ソーシャルタギングのサーベイとしては [6] が非常に良くまとまっている.

5. まとめと今後の課題

本稿では, PLSI を用いて, folksonomy の三つ組データを解析することによる, 自動分類体系構築に関する提案を行った. 提案手法では, ソーシャルブックマークサービスのようなユーザ参加型の folksonomy システムにおいて, タグ間の距離尺度を用いることによって, 分類体系としてふさわしいタグを抽出しそれらを用いて情報発見を容易にする階層構造やグループ分けを行う. 提案手法のメリットとしては, Yahoo やクロスリスティングといった従来のウェブディレクトリと異なり, ユーザのブックマークデータを元に構築するため, 専門家を雇う必要がない点がある.

ソーシャルブックマークサービスは, ボトムアップな集合知形成の典型例として注目を集めており, 今後このサービスから学んだ教訓は次世代のリソース分類技術にも大きく影響を与えていくと予想される. 提案手法は分類対象およびタグの種類についてはいかなる制約も置いていない. 三つ組データからなる Folksonomy システムであれば, 構造的な分類体系の機械的な構築を支援することが可能である. 今後の課題としては, 少人数のブックマーカの動きに対するロバスト性の確保を挙げる.

文 献

[1] Christopher H. Brooks and Nancy Montanez. Improved annotation of the blogosphere via autotagging and hierarchical clustering. In *Proc. WWW*, 2006.

[2] del.icio.us. `http://del.icio.us`.

[3] Paul Heymann and Hector Garcia-Molina. Collaborative Creation of Communal Hierarchical Taxonomies in Social tagging Systems. In *InfoLab Technical Report*, 2006.

[4] Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Proc. International Workshop on Artificial Intelligence and Statistics*, pp. 65–77, 2001.

[5] Clay Shirky. Ontology is Overrated: Categories, Links, and Tags. http://www.shirky.com/writings/ontology_overrated.htm

[6] JakobVo β. Tagging, Folksonomy & Co - Renaissance of Manual Indexing? In *Preprint*, 2007. http://arxiv.org/abs/cs/0701072v2.

[7] Xian Wu, Lei Zhang, and Yong Yu. Exploring Social Annotations for the Semantic Web. In *WWW*, 2006.

[8] 株式会社はてな. はてなブックマーク `http://b.hatena.ne.jp`.

[9] 上田修功, 中野良平. 確定的アニーリング EM アルゴリズム. 電子情報通信学会論文誌 D-II, 第 J80-D-II 巻, pp. 267–276, 1997.