ソーシャルブックマークにおけるタグの派生関係の解析

川中　翔†　佐藤　周行†

† 東京大学 { 基盤情報学専攻, 情報基盤センター }
E-mail: †{kawanaka,schuko}@satolab.itc.u-tokyo.ac.jp

あらまし　我々は従来存在しなかった物や概念が作り出される現象を指して創造と呼ぶ．創造された対象は何らかの新規性を持つと同時に，既存の対象の性質を継承する．このような対象の派生関係は，その系統を理解する上で便利であり，解析すべき重要なものである．本研究では，ソーシャルブックマークにおける概念を記述するタグを解析することで，概念の派生関係を抽出する手法を提案する．提案手法は，ある入力タグがあるとき，出現直後の共起率が高い他のタグが派生関係の候補タグであり，そのうち出現順が早いものが派生関係の親となるタグである，との仮定に基づく．本研究でははてなブックマークを利用した評価実験によって，提案手法の一定の効果を確認した．
キーワード　Web マイニング, Web フィルタリング, 時系列解析, Folksonomy, ソーシャルブックマーク

Extracting Derivation Network of Tags in Social Bookmark

Sho KAWANAKA† and Hiroyuki SATO†

† {Department of Frontier Informatics,Information Technology Center}, the University of Tokyo
E-mail: †{kawanaka,schuko}@satolab.itc.u-tokyo.ac.jp

Abstract　"Creation" is defined as making something exist that has not existed before. Although a created object is characterized as having a kind of newness, it inherits characteristics of its source. This kind of object derivation is useful for understanding dynamism of objects. In this paper, we analyze tags in social bookmark as the target of concept derivation. We propose a method to find the network of concept derivation. Our method is to list tag pairs that show high co-occurred immediately after their emergence. In the evaluation by using Hatena Bookmark, we show effectiveness of our method.
Key words　Web Mining, Web Filtering, Tempora Analysis, Folksonomy, Social Bookmark

1. はじめに

近年，様々な技術の進歩により，情報の伝播性が高まり，また情報の複製が容易となっている．このような情報通信のインフラの整備は，人々の情報の発信性を高め，誰もが自由に情報を発信し，相互に影響しあう時代を迎えつつある．今後はより情報の相互影響性についての理解が必要となり，新しい情報がどのように創られ，どのように後の情報に影響を与えているかなどの，時系列の前後を考慮した情報の関係評価が重要になることが予想される．

我々は従来存在しなかった物や概念が創り出される現象を指して創造と呼ぶ．創造された対象は何らかの新規性を持つと当時に，既存の対象の性質を継承する．一例として，プログラミング言語「Ruby」は，Perl，Java などの既存のプログラミング言語の特徴を組み合わせて誕生した言語であり，後には Ruby を元にしたフレームワーク「Ruby on Rails」が誕生した．このような派生関係の抽出は，集団における文化や知識の系統を理解する上でしばしば重要となる．

しかしながら，様々な概念の出現において、その元となった概念を抽出するタスクは一般にコストが高くつき容易ではない。抽出するには、各概念同士の定義と、その発生過程を定性的に論じるのが自然な手法であり，自動的に抽出することは難しい．すなわち，現状においては，人々がある概念の派生関係を調べたい場合，その概念の記述があるページを閲覧したり，検索エンジンを利用するなどの方法が一般的である，これらは網羅性に弱く，多大な労力を要し，また派生関係は定量性を有していない．

このような状況に対して，本研究では異なるアプローチからの，概念間の派生関係の抽出を試みている．本研究ではその派生関係について，概念の定義などから論じるのではなく，メディア上における概念の「人々による利用データ」を分析することで，近似解析を行う．本研究のアプローチでは，時空間メディア上において，人々が各概念をいつから使い始めたか，また頻繁に一緒に用いられる他の概念は何であるか，などのデータを統計的に解析する手法により概念の派生関係の抽出を試みている．

このような時間的な依存関係を扱う研究をおこなう場合，解析のためのデータに付属する時間情報の取得が困難であるとの問題が存在した．充実した時間情報を含むデータは，一部のデータ所有者を除いて取得できない状況が多く，これまでの時系列的な解析は自己相関分析などプリミティブなものが多くを占める．しかしながら近年登場した「Consumer Generated Media(CGM)」と呼ばれる，エンドユーザが情報を生成するメディアでは属性情報が定型化されており、時間情報が取得しやすくなっている．本研究では，解析対象として CGM の一つであるソーシャルブックマークサービス (SBM) を選択し，利用者によるログデータを分析することで，概念の派生関係の抽出する．SBM では多くのエンドユーザによって意味付けがなされており，またタイムスタンプが自動的に保存される．また SBM は情報の取得性に加え，ユーザ間の相互影響も強いため，情報の相互関係を調査するにあたり適している．我々の手法では，このような SBM の時間情報の取得しやすさとその相互影響を含む集合的な知識を利用し，大量のデータを統計的に解析することで時系列的な概念間の関係の抽出を試みている．

本稿の構成は次の通りである．2 章では派生関係の定義と，解析対象となる SBM について述べる．3 章では派生関係抽出のためのアプローチと，それをユーザに分かり易く示すインターフェイスを提案する．4 章では提案手法を実際のデータに適用した実験について述べる．実験では提案手法をはてなブックマークに適用し，タグ間の派生関係を実際に抽出する．また抽出された派生関係を Wikipedia から抽出した派生関係と比較することで評価を行う．さらに，得られた派生関係をナビゲーションするインターフェイスを実装する．5 章では背景分野について述べ．最後に 6 章ではまとめと今後の課題を述べて締めくくる．

2. 概念の派生関係の定義，ソーシャルブックマークにおける概念

2.1 派生関係の定義

本節では派生関係についての定義を与える．

ある概念があるとき，その概念の成立 (出現) に影響を与えた他の概念がある．本研究では，このような関係を派生関係と呼ぶ．

ここで「派生関係」などの本研究で用いる用語を次のように定義する．ある概念 $(A_1, A_2, ..)$ が成立したあとに，その影響を受けて，他の概念 B が成立したとする．この場合，各 $(A_1, A_2, ..)$ を B の親概念と呼び，B を各 $(A_1, A_2, ..)$ にとっての子概念と呼ぶとする．これらの関係を派生関係と呼ぶこととする．これらの関係は図.1 のように表される．このような派生関係の抽出が本研究の解くべき問題である．

2.2 ソーシャルブックマーク

本研究では Web 上のソーシャルブックマークサービス (SBM) を解析することで研究を行なう．SBM とは Web 上のブックマーク管理・共有サービスのことある．サービス内でユーザは Web 上の様々なドキュメントについてタグを付加することで意味付けを行なっている．(タグとはユーザが任意に決められるフリーワードである．) ユーザはタグを利用することで，ドキュメントが含む概念を記述しており，タグは概念と同一視して考えることができる．またタグ付けが行われた時刻は自動的に保存される．

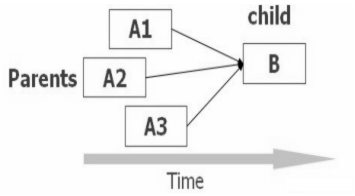

図 1 「派生関係」の定義

SBM における大量のタグ付けの情報は，ユーザ，ドキュメント，タグ，時刻などを含む知識モデルであると考えられ，その性質などの解析がなされている．

SBM を解析する場合の基本性質について次に示す．SBM における基本情報は，ある時刻 t_i において、あるユーザ u_j が、あるタグ c_k を、あるドキュメント d_l に付ける現象で、これは記号的には $Annotation(u_j, c_k, d_l, t_i)$ と表される．本研究で特に時刻について着目し，$Annotation$ を大量に収集し・統計的に分析することで依存関係の抽出をおこなう．

3. ソーシャルブックマークからのタグの派生関係抽出手法

本章ではソーシャルブックマークにおけるタグの派生関係抽出手法を提案する．提案手法は特に時間情報に着目する．

3.1 解析対象，派生関係のモデル，抽出手法

本研究では人々の概念の利用データから，概念の派生関係を抽出する手法を提案する．提案手法では，特に概念の出現時期と，その後の他の概念との共起に着目した解析を行う．

解析対象とする概念の利用データ，抽出する派生関係, 抽出手法の詳細を次に示す．このようなモデルにより，派生関係の抽出について，多くの場面で汎用的に用いることができる．

解析対象となる人々による概念の利用

解析対象とするユーザの概念の利用データは次の通りである．ユーザの集合 $U \in \{u_i..\}$ が存在する．ユーザは (任意の) 現象について概念 $\{c_i, .., cn\} \subset C$ を用いて表すことができ，新たな概念を用いることもできる．本研究ではこのような「ユーザ u_i が，時刻 t において，概念 $c_i, .., cn$ を利用する」という行動: $Action(u, c_i.., t)$ を収集する．図 2 は，$Action$ の例で，初めにユーザ a がタグ A を用い，次にユーザ b が A, B を，最後にユーザ c が A, C を用いている．また，特に本研究では SBM におけるタグを概念と同一に考えて解析をおこなう．SBM を解析対象とした場合 $Annotation$ が $Action$ となる．

派生関係のモデルとタグ間の関係

上記の $Action$ の集合から抽出するタグ間の派生関係を次に示す．派生関係は 2.1 節に述べた定義に基づく．

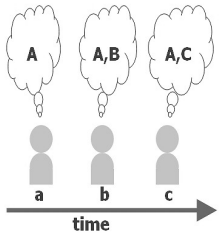

図 2　ユーザの概念利用

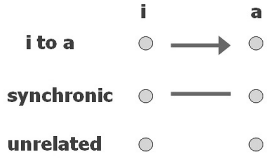

図 3　派生関係のグラフ表示との対応

タグの集合 T があるとき，各タグ a について，他の各タグ i との関係 $relation(a,i)$ を，次の 3 種によって定める．$relation(a,i)$ のとりうる関係は，$unrelated$(関係の弱いタグ同士)，$i\ to\ a$(a は i から派生)，$synchronic$(同時発生) の 3 種である．これらの関係と例示したグラフの表示との対応は図.3 に示す通りである．$syncronic$ はタグの親子関係の前後が明瞭でないものに対応した仕様であり，必要に応じて用いる．このようにタグ間の派生関係を捉えるために，各タグについてその親タグを $Annotation$ の集合から抽出する．

派生関係抽出手法

タグ間の関係を定めるための仮説と手法は次の通りである．

提案手法では概念を表すタグの派生関係を抽出するにあたり，タグの利用における次のような現象が，一般的に起きていると仮定する．ある概念を表すタグ a と，それから派生した概念を表すタグ b があるとき，タグ b が出現した直後にはよくタグ a と共起する（同時にタグ付けがなされる）．また同時期に出現した関係の深い概念のペアがあるとき，それらはタグ付けにおいて近い時期に初出現する．

この現象を元に，次に示す抽出のための仮説を立てる．あるタグ a が出現したときに，その直後に a と共起する（同時にタグ付けされる）各タグ (b) は，a の出現および意味の確立に大きく影響した（a の親）タグである可能性が高い．また各タグ b は A に比べ出現時期が離れていれば，a の親概念である信頼性が高く，また出現時期が近ければ a と同時期に発生した可能性が高い．

上記の仮説を元にした，入力タグ a の親タグを抽出する手法を示す．

（1） a の出現直後 (a がはじめて使われてから,「N 回目に達するまで」や「D 日以内」などの条件を満たす範囲) に a との共起度が高いタグ各 i を M 件取得する．(D,N,M は閾値)

（2） a，i が初めて出現した（初めてタグ付けが行なわれた）日時の差を $emerge(a,i)$(a が早く出現したとき値は負を取る) とすると，a と (1) で選出された各 i の関係は次のように定められる．(X は閾値)

$$relation(a,i) = \begin{cases} i\ to\ a & if\ emerge(a,i) > X \\ syncronic & |X| > if\ emerge(a,i) \\ unrelated & others \end{cases} \quad (1)$$

上記手法によって各タグの親タグを抽出し，さらにそれらを組み合わせることによってタグ空間全体の派生関係が作成される．なお，本手法で用いる共起度については指標 $AEMI$(Augmented Expected Mutual Information) [**?**] を用いる．$AEMI$ は確率を考慮した精細な共起度を測るための指標で以下のように表される．

$$AEMI(a,b) = MI(a,b) + MI(\overline{a},\overline{b}) \quad (2)$$

$$-MI(a,\overline{b}) - MI(\overline{a},b) \quad (3)$$

$$MI(a,b) = P(a,b)log\frac{P(a,b)}{P(a)P(b)} \quad (4)$$

この場合 $P(a)$ はタグ付けにおいてタグ a が用いられる確率であり，$P(a,b)$ はタグ付けにおいてタグ a と b の両方が用いられる確率である．さらに $P(\overline{a})$ はタグ a が投稿において用いられない確率を表す．MI は共起率を評価するための一つの指標であり，$AEMI$ は MI を組み合わせることで，スケールを考慮した確率的な共起度の高さを測ることができる．

また，考慮されるべき現象として，SBM では，一般に synonym と ambiguity という問題が発生する．synonym とは複数のタグが一つの実体を表すことで，ambiguity とは一つのタグが複数の実体を表すことである．これらは，SBM における一般的な問題として，大きなテーマとなっており，広く研究されている．本研究ではこれらの問題の解決を主眼には置かず，これらが解決された状況での提案手法の有効性を議論することとする．

3.2　可視化インターフェイス

本節では，抽出された派生関係をユーザに分かり易く表示するインターフェイスを提案する．

インターフェイスの目的は，概念の派生関係をユーザに提示し，その概念周辺の文化や知識の系統についての理解を支援することである．ここでの支援とは，ユーザにとって少ない労力で，直感的に理解できることを想定している．

インターフェイスでは，予め各タグについて，その派生関係，利用回数，共起の強さ，タグを示すアイコン（画像）をデータベースに格納する．ユーザから任意のタグについて問い合わせがあると，図.4 に示されてるように関係のあるタグの一覧をネットワークグラフによって示す．各タグを表すノードは関係の性質を表すリンクによって結ばれ，そのリンクは共起に強さにより太さを変えて表示される．また各タグ表すノードはそのタグを表すアイコンによって表示され，その大きさはタグの利用頻度を表す．また親概念のみ，もしくは子概念のみの抽出も可能とすることで柔軟な表示への対応を可能とする．

このようにすることで各タグ間の系統的な関係について，視覚的に分かり易く提示する．

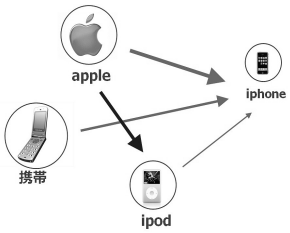

図 4　可視化インターフェイス

4. はてなブックマークを解析対象とした提案手法の評価実験

提案手法を実際の SBM のデータに適用した実験をおこなった．目的は手法の有効性を測るためである．実験では提案手法をはてなブックマークのタグ時空間に適用し派生関係を抽出した．また得られた派生関係の評価をするために，Wikipedia から手動で抽出した派生関係との比較を行った．

4.1　データセット

今回データセットとして用いたのは 2008 年 10 月中に取得した，はてなブックマークのデータである．はてなブックマークは SBM の一つで，各ページのタグ付けの情報が RSS 形式で提供されている．過去に遡ってデータの取得が可能であり，本研究に適しているため選択した．はてなブックマークは，外部のドキュメント等にリンクを貼るタイプの一般的な SBM であり，ユーザは体系的に情報を扱うことを志向している．はてなブックマークにおいてタグ付けを行なう時には，既存の利用頻度の高いタグが表示されるため，者の影響があると考えられる．取得したデータの詳細は表.1 に示す通りである．

表 1　取得したデータ量

オブジェクト数	928421
ユニークなタグ数	241331
アノテーション数	16307967

今回取得したデータでは，初めてタグ付けがなされたのは 2005 年 2 月で，2005 年の総ポスト数 (複数のアノテーションを同時に含む) は 248432，2006 年は 72957，2007 年は 66132，2008 年は 77399 となっている．

4.2　親タグの抽出例

本節、まず例としてタグ「ニコニコ動画」とタグ「Web」の共起関係について示し，それぞれのタグの親タグを抽出する．「ニコニコ動画」は 2006 年 7 月に初めて出現し，その後急速に利用数を伸ばしたタグ (使用回数 37 位) で，サービス開始当初には存在しなかったタグであると考えられる．一方で「Web」は 2005 年 2 月に初めて出現し，使用回数は 2 位であり，サービス開始よりよく使われていたタグであるといえる．

これらのタグについて閾値 N=200, M=10(予備実験結果から適切な値として設定) として，出現直後の共起度が高いタグを抽出したものを表.2 に示す．また比較対象として M=∞として，アノテーション空間全体で共起度が高いタグを共に示している．タグ「ニコニコ動画」については，M=∞では「vocaloid」(2007 年 6 月登場),「初音ミク」(2007 年 8 月登場)，などタグ

表 2　ニコニコ動画,Web とそれぞれ共起度の高いタグ

ニコ..(N=200)	ニコ..(∞)	Web(200)	Web(∞)
動画	初音ミク	portal	design
動画共有	著作権	design	tool
youtube	動画	news	blog
2ch	ネタ	tool	css
ひろゆき	vocaloid	tips	service
エロゲ	music	便利サイト	tips
2 ちゃんねる	これはすごい	blog	javascript
ネタ	アイドルマ..	css	google
web サービス	音楽	ホームページ	ネタ
niconico	idolm@ster	.htaccess	デザイン

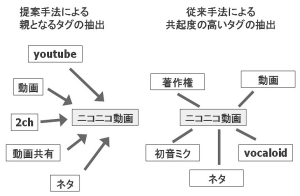

図 5 「ニコニコ動画」の派生関係

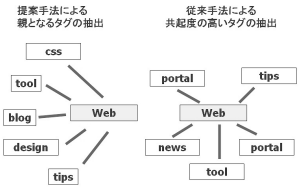

図 6 「Web」の派生関係

「ニコニコ動画」出現時には存在しなかったタグが登場しており，M=200 と異なるタイプのタグが抽出されていることが分かる．

これらのタグとの初登場時期を比較 (閾値 X=20) し，関係を抽出すると図.5, 図.6 のように表される．ニコニコ動画の場合は図中の左側の提案手法のグラフで示されているように，タグ同士の前後関係が明確になっているが，Web の場合は従来手法の方と提案手法の違いが分かりにくくなっている．

4.3　派生関係グラフの統合

次に各タグについて派生関係を抽出し，それらを複数組み合わせたタグ (部分) 空間の派生関係を抽出する．図 7 は，入力ノード数:50，一入力ノードあたりの親タグ数 (M):5，出現直後を表す閾値 (N):200，同時発生の判定の閾値 (X):0 として，50 のノードの派生関係を組み合わせたグラフである．図の上側ほど新しいタグで，下側ほど古いタグである．ノード数は 92，リンク数は 201 で リンク数/ノード数(密度) は 2.18 である．このグラフでは X を 0 としているため，同時発生は存在せずグラフは単方向で表される．X の値を大きくとるほど，グラフは双方向性を持ち，形状も円形に近づく．

図 8, は作成したグラフの一部分を切り取ったもので, ニコニコ動画などのというタグの周りが抽出されており，それらと関係の深いタグが抽出されている．この例では直感的には妥当な派生関係が作成されている．

つづいてネットワークの特徴を捉えるために，入力ノード数と，一入力あたりの親タグ数を変化させたときの密度の変化を調べた．図 9 は入力ノード数を変化させた場合の密度の変化で,

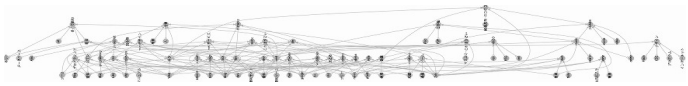

図 7　統合された派生関係 (上側が新しいタグ)

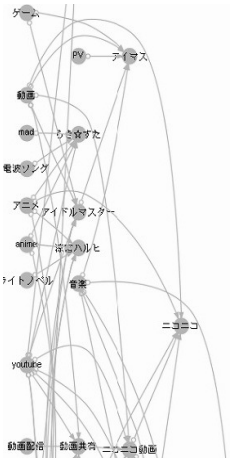

図 8　統合された派生関係 (拡大図)

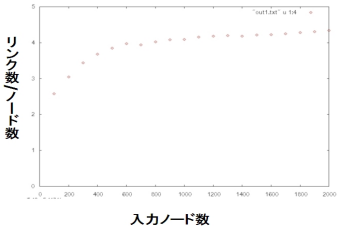

図 9　ネットワークの密度の変化 (入力ノード数:x, 親タグ数:5)

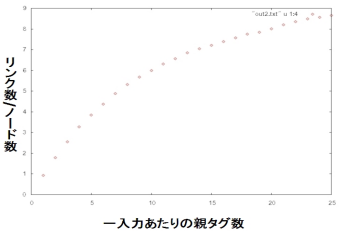

図 10　ネットワークの密度の変化 (入力ノード数:500, 親タグ数:x)

図 10 は親タグ数を変化させた場合の密度の変化である．パラメータとネットワークの構造の関係は定式化されていないが，その値の変化をみることである程度意図するサイズと密度のネットワークを作成することができる．

さらに，各ノードの out リンク数の次数分布を調査した．out リンクが多いノードほど多くのタグの親になっているノードである．結果として図 11 のように，パラメータを変化 (時間情報を利用しない，入力ノード数や親タグ数を変えるなど) させても次数の分布に大差はなかった．図中の直線は $y = 200 * x^{-0.8}$ で表される式を示したものである．はてなブックマークの共起を元にした，提案手法による out リンク数の分布は $\gamma = 0.8$ 前後のべき乗則として固有的に表されると考えられる．

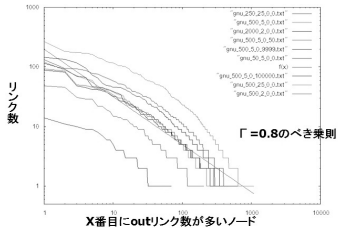

図 11　out リンクの次数分布

4.4　Wikipedia との比較による派生関係の検証

4.4.1　概　　要

提案手法で得られた派生関係の評価を行うため，Wikipedia [10] から取得した派生関係との比較実験を行った．解析に利用する Wikipedia 日本語版は 2009 年 1 月現在で 557123 の記事が存在し，一定の信頼性ある語彙を有する比較対象として利用する．

比較実験の目的は提案手法の効果性，特徴を評価するためである．実験のプロセスは次の通りである．

（1）　入力タグを α 個用意．

（2）　提案手法と比較用それぞれの親タグリストの取得．

- (1) の α 個のタグについて提案手法によって，それぞれの親タグリスト (T) を取得する．
- (1) の α 個のタグについて，wikipedia のそのタグ (単語) に対する記述から，親タグリスト (W) を取得する
- (1) の α 個のタグについて，時間情報を用いない単純共起率によって，親タグのリスト (T') を取得する．

（3）　検証．

- T と W を比較し，その一致度 (タグ同士の一致率および W を正解としたときの再現率) を検証する．
- T' と W を比較し，その一致度 (タグ同士の一致率および W を正解としたときの再現率) を検証する．
- また T と入力タグの，出現時期の差をもとめ，時間の前後の信頼性について検証する．

なお Wikpedia からの派生関係の抽出については次のプロセスでおこなう．まず入力タグについて記述されたページから，リンクが貼られている単語のリストをまず取得する．次に単語のリストの中から親概念 (単語) として有力な単語を手動で取得する．

4.4.2　親タグの抽出と，その一致率および再現率

入力ノードリストは下記の手続きによって $\alpha = 20$ として選出した次の 2 種類のリストである．

- 2006 年以前に出現した 10 個のタグ
- 2006 年以降に出現した 10 個のタグ

サービス開始時から存在したタグと，サービス開始語に出現したタグを区別するため，それぞれのタグを 10 個ずつ入力タグとして設定した．

選出の処理として，まず，後者について 2006 年以後に出現したタグのうち利用頻度上位 10 個のタグを抽出した．(Wikipedia にページが存在するものに限る．) 次にそれら 10 個のタグのジャンルと近いジャンルのタグを 10 個 2006 年以前の出現タグから選出した．

それぞれの入力タグリストは表 3 に示す通りである．(なお

表 3　入力タグ

2006 年以前	ジャンル	2006 年以後	ジャンル
mixi(28461)	サイト名	ニコ..(56965)	サービス
しょこたん (549)	人物・キャ..	twitter(27758)	サービス
flash(68801)	製品	初音ミク (15141)	製品・キャ..
excel(8111)	製品	gigazine(12947)	ニュースサイト
podcast(5964)	サービス	vocaloid(12662)	技術
sns(28112)	サービス	idolm@..(12346)	タイトル
p2p(6521)	技術	涼宮ハルヒ (5538)	タイ..・キャ..
エヴァ(1530)	タイトル	池田信夫 (4604)	人物
デスノート (232)	タイトル	らき☆すた (4508)	タイトル
森博嗣 (546)	人物	iphone(18308)	製品

実験では，はてなブックマークと Wikipedia の語彙の違いを考慮し，Wikipedia に存在しないタグについては予め排除した．また Synonym を考慮し Synonym 同士のタグについては統一した．）

さらに，これらのタグについて親タグ T,W をそれぞれの手法によって取得した．(なお，提案手法のパラメータは $M = 10, N = 200, X = 50$ で，同時発生と親タグを両方取得した．）表 3 のタグについて，Wikipedia から抽出された親タグのリストを表 4 に示す．

また，提案手法によって抽出された親タグのリストと，それらと Wikipedia から抽出されたタグとの一致判定を示すのが表 7 である．(提案手法によって取得されたタグが，Wikipedia から抽出されたタグと一致していたら○，部分的に意味を表していたら△，一致していない場合は×としている．）

表 5 は，一致判定と再現率の全体結果をまとめたものである．提案手法によって取得したタグ T50 件が W(40 件, 41 件) と一致した確率は 25%弱で，T100 件が W と一致した確率は 15%強であった．また W の再現率は 40%強であった．これは本手法によって抽出した親タグ合計上位 100 個 $(T, 20*5)$ のうち，25%弱が 81 個 (40+41) の Wikipedia から抽出した親タグ (W) と一致し，T を 200 個にした場合は 15%となることを示している．また T を 200 個にしたとき、W81 個のうちの 40%強以上まで抽出されたことを示している．

表 6 は，単純共起率によって抽出された親タグリストと，Wikipedia から取得した親タグリストの一致度の全体結果をまとめたものである．これは単純共起率によって抽出した親タグ合計上位 100 個 $(T', 20*5)$ のうち，20%弱が 81 個 (40+41) の Wikipedia から抽出した親タグ (W) と一致し，T' を 200 個にした場合は 15%弱となることを示している．また T を 200 個にしたとき、W81 個のうちの 40%近くまで抽出されたことを示している．

4.4.3　親タグの一致率および再現率からの考察

表 5 によると，結果として提案手法によって抽出された親タグ (T) と，Wikipedia によって抽出された親タグ (W) との一致率は 25 % (上位 50 件),15 % (上位 100 件) で T による W の再現率は 40 %強であった．

これらの指標が示す値は絶対的に高いものではないが提案手法によって抽出されたタグと Wikipedia という信頼のあるに比較対象おける単語との，一定の一致性はみられた．一方で相違点も多々みられた．タグの一致がみられなかった主要因として両空間 (Wikipedia とはてなブックマーク) における語彙の違いが挙げられる．Wikipedia では, はてなブックマークのタグと比較して，比較的ある概念の，制作者情報や元々のジャンルなどが単一的に記載されていることが多く，様々な抽象度によって表されたジャンルや，その概念が影響を受けた他の概念が記載されていることが少ない．

結果として Wikipedia における制作者情報 (GAINAX, ニワンゴ, Obvuous, クリプトン・ヒューフューチャーメディアなど) の制作者情報は多くが提案手法によっては抽出されていない．

このようにそれぞれのメディアは異なる特徴の語彙を有しており，それらが表す親タグの特徴も異なっている．

より定型化された堅い背景情報を取得するなら Wikipedia を利用することが望ましく，人々の利用に裏打ちされた，様々な抽象度で表された緩い親タグや，間接的な影響元となった親タグを取得するなら，Folksonomy からの提案手法による抽出が向いていると考えられる．

これらは抽出されたタグリストを俯瞰したことからの推論であり，定量的に分析することは今後の課題である．

4.4.4　提案手法と単純共起率による結果の違い

表 5 と表 6 を比較したとき，2006 年以降のタグの方が，2006 年以前のタグに比べて精度，再現率がともに 30%に落ち込んでいることが分かる．この要因としては，単純共起率による親タグ (T') の抽出では入力タグより明らかに遅い時期に発生したタグが取得されてしまい，他の有力な入力タグが親タグ候補が抜けてしまうことが主な原因であると考えられる．

表 4　Wikipedia から抽出された親タグのリスト

タグ名	親タグ (Wikipedia)
mixi	ソーシャル・ネットワーキング・サービス,Web
しょこたん	アイドル, タレント, 漫画家, 声優, アニメ
flash	動画, ゲーム, ソフトウェア,Web
excel	ウィンドウズ, マック, 表計算ソフト, アプリケーション
podcast	オーディオ, ビデオ, ウェブログ,ipod,broadcaset,mp3
sns	社会的ネットワーク,web
p2p	コンピュータネットワーク, 分散コンピューティング アプリケーション
エヴァ	アニメ,GAINAX,SF, 貞本義行, 漫画
デスノート	漫画, 週刊少年ジャンプ, ダーク・ファンタジー
森博嗣	小説家, 推理作家, 研究者
ニコニコ動画	ニワンゴ, 動画,youtube, 2ちゃんねる, サービス
twitter	ブログ, チャット,IM(インスタントメッセンジャー)
初音ミク	クリプトン・フューチャーメディア, ヤマハ, 音声合成, デスクトップミュージック, ソフトウェア, キャラクター, VOCALOID, 声優, 藤田咲
gigazine	ブログ, ニュースサイト, アニメ
vocaloid	ヤマハ, デスクトップミュージック, 音声合成, ソフトウェア, 声優
idolm@ster	ナムコ, アイドル, アーケード, シミュレーション
涼宮ハルヒ	谷川流, ライトノベル, いとうのいぢ, テレビアニメ, 学園,SF, ラブコメ, 京都アニメーション
池田信夫	経済学者, メディア, 自由主義, 情報通信, 著作権
らき☆すた	美水かがみ, 4コマ漫画, ゲーム・アニメ
iphone	アップル, スマートフォン, タッチパネル,ipod

表 5　Wikipedia タグとの比較による一致率と再現率

タグ	5 件一致率	10 件一致率	再現率
2006 年以前	12/50	16/100	17/40
2006 年以降	11.5 /50	17.5/100	17.5/41

表 6　Wikipedia タグとの比較による一致率と再現率 (単純共起率)

タグ	5 件一致率	10 件一致率	再現率
2006 年以前	12/50	16/100	16/40
2006 年以降	8.5 /50	13/100	13/41

表 7　提案手法によって抽出された親タグおよびその一致性

タグ名	親タグ (提案手法)
mixi	○ SNS, × blog, ○ web, × tool, × Firefox, × software, × tools, × tips, × interview, × bookmarklet
しょこたん	○アイドル, × blog, ×ネタ, × 2ch, × tv ×動画, ○声優, ×コスプレ, × video, ○芸能人
flash	○ game, × ajax, × javascript, ○ web, × actionscript ○ movie, ×ネタ, × site, × flickr, × design
excel	× tips, × office, × web, ○ windows, × wiki × ajax, × tool, × html, × java, × web2.0, × web サービス
podcast	○ music, ○ ipod, × radio, × software, × tv × RSS, ×はてな, × media, × hatena, ○ blog
sns	× mixi, × blog, ○ web, × hatena, × photo × music, ×携帯, × community, × video, × news
p2p	× firefox, × winny, ○ software, × security, × google × book, × skype, × music, × web2.0, × web
エヴァ	○アニメ, × 2ch, ×ネタ, ×動画, ×オタク × blog, × neta, ×考察, ×まとめ, × web
デスノート	×ネタ, ○漫画, × 2ch, ×映画, ×ニュース ×猫, × youtube, ×*anime, × movie, ×サザエさん
森博嗣	× blog, ×押井守, × book, ×社会, × life ×ブログ, × anime, ×アニメ, ×本, ○作家
ニコニコ動画	○動画, ○ youtube, ○ 2ch, ×ネタ, △動画共有 △ web サービス, ×ひろゆき, ×エロゲ, ×まとめ, ×これはすごい
twitter	×まとめ, × mobile, × webservice, × tool, × web × greasemonkey, × sns, × software, ○ blogs, △ irc
初音ミク	×ニコニコ動画, ○ vocaloid, △音楽, ○声優×まとめ ○ dtm, ×動画, ○ software, × 2ch, ×萌え
gigazine	× software, × design, × webservice, × free, × web × photo, × firefox, × windows, × material, ○ blog
vocaloid	× niconico, △ music, ○ software, ○声優, ×ネタ × interview, × 2ch, ○ dtm, ×まとめ, ×同人
idolm@ster	× youtube, △ game, ×ネタ, × movie, ×まとめ × xbox360, × interview, × anime, ×電波ソング, × MAD
涼宮ハルヒ	○アニメ, ×ネタ, ○ライトノベル, × 2ch, ×まとめ △ book, × youtube, ×声優,flash, ×公式, ×動画
池田信夫	× blog, × RDF, × TrackBack, × ping, × DC ○著作権, ○経済, ×社会, × web2.0, × long tail
らき☆すた	× niconico, △ anime, × youtube, × mad, ×電波ソング ×涼宮ハルヒ, ×まとめ, ○ 4 コマ漫画, ×ネタ, ×動画
iphone	○ apple, △ mobile, ○ ipod, × mac, × gadget ×新製品, ×携帯電話, × news, × youtube, × presentation

表 8　親タグとの出現時期の差

タグ	0 以下	1〜30	31〜100	101〜
2006 年以前	18	54	3	25
2006 年以降	2	0	2	96

具体例として，T' では vocaloid というタグを入力とした場合に，初音ミク，鏡音リン，鏡音レンなどの明らかに後に出現したタグが抽出されてしまっている．

この結果は，時間情報を考慮することで，入力タグの種類によっては，より精度の高い親タグのリストを抽出できることを示している．

4.4.5　親タグと入力タグとの出現時期の差

提案手法によって取得した親タグの，入力タグとの出現時期の差を調査した．入力タグとの差が大きいほど同時発生ではなく，親タグである信頼性が高い．表 8 は入力タグと親タグとの出現時期の差の分布をまとめたものである．親タグとの差が大きいのは 2006 年以後のタグである．2006 年以前のタグはこのように同時発生である可能性が高くなり，2006 年以降のタグは親タグである信頼性が高くなる．このように，タグ同士の時間的な差が明瞭になるのは，サービス開始以後に出現したと考えられるタグ間の関係である．

4.5　可視化インターフェイスの実装

抽出した派生関係を提案インターフェイスにより，ユーザに分かり易く示すナビゲーションシステムを実装した．(各タグを表す画像については Google Image から取得した．システムの実装については Flash と Javascript によっておこなった.)

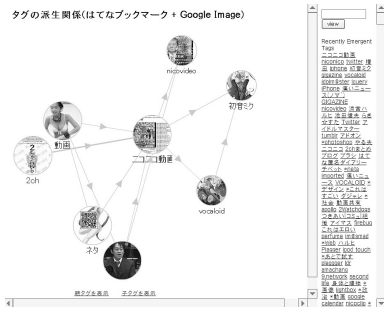

図 12　「ニコニコ動画」の派生関係 (インターフェイス)

図 12, がシステムの基本画面で，タグを選びクリックすると，そのタグについての派生関係が表示される．また下のボタン (親タグ，子タグの表示) をクリックすることで，そのタグの親タグや子タグのみの表示も可能となる．

なお，システムは Web (http://c4f7.cc.u-tokyo.ac.jp/KAWANAKA.Sho/flash/) において公開しているので詳細はそちらを参照されたい．

4.6　考　　察

4.6.1　タグの新規出現と SBM のサービス開始時期

提案手法では，各タグが初めて使われた日時を，タグの前後関係を決定するパラメータとして利用している．しかしながら，SBM が開始以前から成立していたタグも多数存在し，それらのタグについては SBM 内における初出現時期は一律に初期に集中し，前後関係を定めるにあたり，日時パラメータの意味が薄れる．実験ではサービス開始時に存在していたタグと，そうでないタグを区別して比較し，後者の方がより効果的に提案手法を用いることができた．SBM の運営期間が長くなるほど，後者の割合が高くなり，より提案手法が有効になる．また将来的に Web における統合的なタグの使用状況についてのデータが充実すれば，より本質外の問題となる．

4.6.2　インターフェイス

本研究では派生関係をユーザに示すインターフェイスに Flash と Javascript を用いた．

既存のタグの利用状況を示すインターフェイスとしてはタグクラウドが近年広く普及した重要な存在である．タグクラウドはテキストベースによってタグの一覧を示し，文字の大きさによって重要度を区別している．複雑な関係性は表していないが，テキストベースで扱える手軽さから広く普及している．

提案手法の派生関係を表すなら，次のように場面に応じて適当な方法を用いるのが望ましい．テキストベースで用いるなら親タグと子タグによって配置を分ける表示法，ブログパーツ的にコンテンツの一部分として用いるなら，小さな画面に対応した軽いパーツの仕様を構築する必要がある．今回提案したインターフェイスは，画面全体で用いる場合のナビゲーションに適している．

4.7 今後の課題

4.7.1 派生関係の妥当性，ネットワークの性質の評価

抽出した派生関係の評価は今後さらに行う必要がある．比較するサンプル数の大規模化，タグを分類した上での種類毎の比較，他情報源から抽出した派生関係との比較などが，考えられる．また被験者による，直感からの評価もまた有力である．さらに利用者が明示的に派生関係を入力できるようなシステムを構築し，そこでのフィードバックを取得することで，派生関係についての理解を深める方法も考えられる．

また派生関係ネットワークの特徴をさらに評価する必要がある．単一方向の有向ネットワークに対応した指標を用いた特徴評価が望ましく，今後の重要な課題である．

4.7.2 タグの変容

提案手法では派生関係におけるタグをネットワークグラフ上の一つのノードとして表し，その親概念は解析を行う時期にかかわらず，一意に定まる．しかしながら，タグの意味は時間と共に少しづつ変化し，タグの派生関係を固定的に表すことは不適切なのではないか，という指摘も考えられる．それに対し，提案手法によって表される関係は，タグの出現時点に主眼を置いた，派生関係であり，その「発生過程」に着目している．一方で，タグが出現しまとまった意味を持った以後に，少しずつ意味をマイナーチェンジし，他のタグとの影響関係が変化する現象も考えられる．このような出現以後の「影響関係」を捉えるには別の解析手法を用いる必要がある．

5. 背景分野

本研究の背景分野として，Web マイニング，Web フィルタリング, 時系列的解析が挙げられ，Web 上における派生関係も注目されている．また解析対象のプラットフォームである Folksonomy(ソーシャルブックマークの特徴的機能) について近年活発に研究がなされており，その分野は上記の分野にまたがっている．

本研究は Web におけるタグ付けというユーザの行動から，タグ同士の関係を発見することを目的としているので，Web マイニングのうち，Web 利用マイニング [1] に該当する．既存の Web 利用マイニングの技術としては Amazon.com [7] などで利用されている協調フィルタリングが広く知られている．

Web フィルタリングは，Web の膨大なリソースから，必要な情報をユーザに便利に提供するための技術の総称である. [4] エンドユーザのタグによる分類は Folksonomy とよばれ，タグ間の関係整理を目的とした研究も行なわれている．本研究では特に派生関係に着目し，より便利なタグの利用やページのフィルタリングに役立てることを目的としている．

時間情報を用いた既存研究は次のように大別される．情報の伝播性に着目した研究，情報のバースト性に着目した研究，重要度決定のための指標として利用する研究，パターン認識のパラメータとして用いる研究である．本研究では特に情報の新規発生に着目し，その派生関係を抽出するために時間情報を用いる．その点で，時間情報と利用しているという共通点はあるが，得ようとする関係が異なるという意味で，上記の研究とはテーマが異なっている．

Web 上の派生関係として「マッシュアップ」という用語が近年普及している. [5] [6] Web サービスや，動画共有サイトやソーシャルネットワーク上の概念のマッシュアップなどが代表的なものであり，これらは一種の派生関係である．またマッシュアップを機能に取り込んだサービスとしてニコニココモンズ [9] が挙げられる．ニコニココモンズでは，システム側が派生関係を明示的に利用し，コミュニティの活性化や情報の便利なナビゲーションの実現を図っている．

Folksonomy は近年出現した重要な知識プラットフォームとして注目を集め，様々な視点から研究がなされている．Golder ら [2] は SBM における，ユーザ，タグ，ブックマークの性質について，ユーザの行動，タグの種類，ブックマークの人気に関する興味深い法則を発見した．丹羽ら [4] は SBM におけるユーザベースの共起度とドキュメントベースの共起度を比較することで，Synonym を高い精度で発見する手法を提案した．Mika [3] は SBM におけるユーザとタグとドキュメントの関係を 3 部グラフの一種として定義し，そこからオントロジーを構築する可能性を示した．

6. おわりに

本稿では，概念の派生関係を定義し，その抽出手法と可視化するインターフェイスを提案した．実験では提案手法を，はてなブックマークのタグに適用し，派生関係の抽出した．また，そのネットワークの性質を密度と次数分布から議論した．派生関係の評価のために，タグの派生関係について Wikipedia との比較によって検証を行い，それらの一定の一致性を確認したが，語彙の違いからの相違点も多々みられた．また特に SBM サービス開始以後のタグについて時間情報が有効に作用することが確認された．さらに提案インターフェイスを元にしたタグナビゲーションシステムを構築した．

今後は派生関係ネットワークの性質解析指標の導入や，より精細な評価が重要な課題である．またインターフェイスの評価，手法の他サービスへの適用，派生関係を機能化したサービスのデータの解析なども課題である．

文　献

[1] Bing, L “Web Data Mining,” Springer-Verlag Berlin Heidelberg (2007)

[2] Golder, S.A. HUberman, B.A. “The Structure of Collaborative Tagging System,” Information Dynamics Laboratory, HP Labs (2005)

[3] Mika, P. “Ontologies are us: A unified model of social networks and semantics,” 4th International Semantic Web Conference (2005)

[4] 丹羽智史, 土肥拓生, 本位田真一, “Folksonomy マイニングに基づく Web ページ推薦システム,” 情報処理学会論文誌 (2006)

[5] IT 用語辞典, マッシュアップとは 【mash up】 : 意味・解説 http://e-words.jp/

[6] IT 用語辞典バイナリ, マッシュアップとは (mashup) http://www.sophia-it.com/

[7] Amazon.com http://amazon.com/

[8] はてなブックマーク. http://b.hatena.ne.jp/

[9] ニコニココモンズ http://www.niconicommons.jp/

[10] Wikipedia http://ja.wikipedia.org