——セマンティックウェブとオントロジー研究会——

Folksonomy の 3 部グラフ構造を利用したタグクラスタリング

丹羽　智史[†a)]　　土肥　拓生[†b)]　　本位田真一[†,††c)]

Tag Clustering Method based on Folksonomy Tri-Partite Graph Analysis

Satoshi NIWA[†a)], Takuo DOI[†b)], and Shinichi HONIDEN[†,††c)]

Abstract.
近年新しいドキュメント分類の形態として Folksonomy(フォクソノミー) が注目されている.Folksonomy は従来の Taxonomy とは対照的に, エンドユーザによるボトムアップでフラットな分類を実現する.Folksonomy は分類速度, 分類の実用性, 分類の適応性など多くの面で優れており, 各種 Web Filtering 技術に応用可能な大きなポテンシャルを有しているのにも関わらず, 現在のところ十分に活用されていない. 本論文において我々は Folksonomy の性質を分析し,Folksonomy の 3 部グラフ構造を利用することで Synonym 問題など Folksonomy 固有の問題を解決しながら効率的に分類データを構造化する手法を提案し,Folksonomy の応用への道を開く.

Keywords. Folksonomy,Taxonomy,Web マイニング,Web フィルタリング

1. はじめに

近年 Web で生み出される情報は日々増加の一途をたどり, また情報ソースも従来のニュースサイトなどに加えブログ,SNS などと多様化がすすんでいる. 我々は限られた時間の中でより効率的に Web の情報を収集, 処理する必要にせまられている.

Web の膨大なドキュメントの中から自分が必要としている情報を取捨選択してくれる技術を Web フィルタリング技術と呼ぶ.Web フィルタリング技術は主にコンテンツ評価機能, コンテンツ分類機能の 2 つの機能を有する (図 1).

図 2 は現在広く普及している Web フィルタリング技術を独自の基準で分類した表である.

Google [10] などの検索エンジンはドキュメント間のハイパーリンク構造を解析し, ページランク [2] を求めることでドキュメントの評価を行っている. またドキュメント内に現れるキーワードをもとに分類を行っている.RSS リーダーなどのアグリゲーターは Web 上で公開されている RSS フィードや ATOM フィードを定期的に巡回し, ドキュメントの更新を確認する. ページランクなどのコンテンツ評価の仕組みは無いが, 検索エンジンよりも新しい情報を集めることができる.

本論文では第 3 の項目であるソーシャルニュースサイトに注目したいと思う digg [8] や del.icio.us [9] はエ

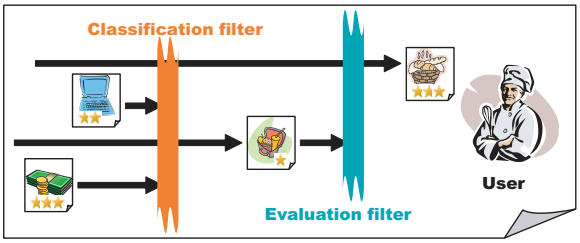


図 1　Web Content Filtering

	Search engine	Aggregator	Social news site
Examples	Google YAHOO! search	goo RSSリーダー	del.icio.us digg
Contents evaluation	○ web-link analysis	× none	○ end-user voting
Contents classification	keyword matching	keyword matching	end-user classification Folksonomy
Contents newness	× a few days	○ a few hours	○ a few hours
Contents domain	◎ Any contents on web	△ Contents with RSS (or Atom) feeds	? Any contents notified by end users

図 2　Modern Web Filtering Services

† 東京大学
†† 国立情報学研究所
a) E-mail: niwa@nii.ac.jp
b) E-mail: tdoi@nii.ac.jp
c) E-mail: honiden@nii.ac.jp

ンドユーザによるドキュメントクリッピングサービスである. サービスごとに仕組みは若干異なるが, ユーザに人気のあるドキュメントにはポイントが溜まっていき, ドキュメントの分類もユーザ自身が行う点は共通している. したがってドキュメント評価もドキュメント分類も完全にエンドユーザに依存しているのが特徴である.

del.icio.us などで用いられているユーザによるドキュメント分類の仕組みは Folksonomy と呼ばれる [5].Folksonomy は分類の速さ, 分類結果の実用性などが非常に優れており,Web Filtering の各種技術へ応用できる高いポテンシャルを有しているが, 現状では十分にそのポテンシャルが活かされていないように感じる.

我々は過去の研究において Folksonomy のデータを利用することで高精度でユーザの趣味嗜好に一致した Web ドキュメントを推薦する Web Recommneder を構築し,Folksonomy の有用性を示した [1].

本論文では Folksonomy の性質を分析し,Folksonomy の 3 部グラフ構造を利用したタグのクラスタリング手法を提案することで Folksonomy の各種技術へのさらなる応用の道を開く.

本論文の構成は以下のようになっている.2 章では Folksonomy の基本的な仕組み, 利点及び問題点について述べる.3 章では Folksonomy の 3 部グラフ構造に基づき Folksonomy の性質に関する仮説を立てる.4 章では仮説をもとに Folksonomy のタグをクラスタリングする手法を提案する.5 章では実験によって仮説を検証する.

2. Folksonomy(フォクソノミー)

2.1 Folksonomy とは

Folksonomy は別名 Social Tagging(ソーシャル・タギング) とも呼ばれ, 伝統的な分類手法である Taxonomy と対比されることが多い. 図 3 は Taxonomy と Folksonomy の分類構造を図示している.

Taxonomy は少数の権威, 担当者があらかじめ分類ヒエラルキーを構築しておき, その後に個々の分類被対象物（今回の場合は文書）をヒエラルキーの最下層に組みいれていくトップダウンな分類手法である. 分類構造の構築フェーズと分類フェーズが分かれているのが特徴である.

Folksonomy はこれとは対象的にボトムアップな分類手法である.Folksonomy の分類構造は個々のエンド

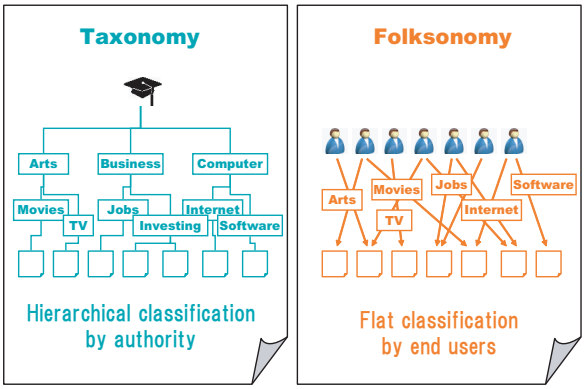


図 3 Taxonomy and Folksonomy

ユーザがそれぞれの文書に ”タグ付け ”を行うことによって構築されていく. タグ付けとはユーザが対象文書の特徴を端的に表すキーワードを自分の主観によって選択する行為であり, 任意のキーワードを文書にタグ付けすることができる. また 1 つの文書に複数のタグを付けることもできる.Folksonomy においては分類フェーズがそのまま分類構造の構築フェーズを兼ねる. 加えて,Folksonomy の分類構造はフラットである.

2.2 Social Bookmark Service(ソーシャルブックマークサービス)

Social Bookmark Service(以下 SBS) はオンライン上で Web ページのブックマークを管理・共有するための Web ベースのサービスである. 個々のユーザは今まで自分のブラウザで管理していたのと同じように SBS 上でブックマークを管理することができる. また SBS 上では登録している全てのユーザのブックマークが公開されているので, 自分と趣味嗜好が似ている他ユーザのブックマークを参考に情報収集に役立てることができる.

SBS 上でのブックマーク文書の分類には Folksonomy が用いられていることが多い. 多くの SBS ではブックマークを登録する際に同時にタグ入力を求められる. 本節で SBS を紹介したのは,SBS が現状で最も大規模な Folksonomy の使用例となっているためである. 現在世界最大の SBS である del.icio.us では 500,000 人以上の会員のブックマークデータとそれに付随するタグのデータが公開されている（図 4）. 本論文では SBS 上のデータを用いて Folksonomy の性質を分析する.

2.3 Folksonomy を扱う上での難しさ

Folksonomy のタグデータを処理する際には, 以下

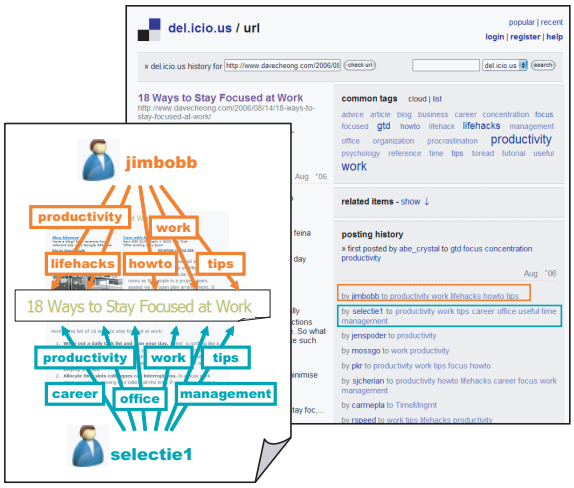


図 4　Folksonomy in del.icio.us

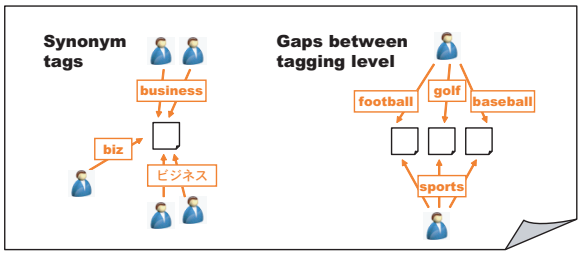


図 5　Difficulty in Dealing with Folksonomy

の点に注意する必要がある.

（1） **Synonym(シノニム) の問題**

Folksonomy ではタグに任意のキーワードを選択できるため, 同じ意味の異なるタグが混在する事態が頻繁に起こる（図 5 左）.Synonym なタグはデータの冗長性をもたらすので, 可能な限り同種タグとしてまとめてしまうのが望ましい. 本論文では 2 つのタグが Synonym か否かを高確率で判定するための手法を提案する.

（2） **タグの分類抽象度の問題**

Folksonomy では個々のユーザが自分に使い勝手を基準にタグを選択するため, さまざまな抽象度のタグが混在してしまう (図 5 右). 本論文ではタグのクラスタリングによる構造化によりこの問題の解決を試みる.

2.4　Folksonomy の特長

Folksonomy は既存の Taxonomy と比較して以下のような特長を有する

（1） **分類が速い**

Folksonomy では何百〜何万人ものユーザが分類を行うため分類速度がとても速い. とくに新しい記事に対する分類速度に強みをもつ.

（2） **分類が実用的**

Folksonomy では各々のユーザが自分自身にとって最も使い勝手がよくなるように文書のタグ付けを行う. 結果として多くのエンドユーザにとって実用的な分類がなされる.

（3） **分類構造が動的に変化する**

Folksonomy ではユーザの分類行動を反映して分類構造が動的に変わっていく. そのためその時代のパラダイムやユーザの意識に適応した分類構造が自動的に構築されていく.

これらの特長を見ると Folksonomy は Document Classification, Recommendation, Web Personalization など多くの分野に応用できる非常に大きなポテンシャルを持っているといえる. しかし現時点では Folksonomy は例えばタグによる文書検索（タグサーチ）など非常に限定された方法でしか利用されていない.

Folksonomy の技術応用を実現するために Folksonomy の性質を分析し,Synonym 問題などを解決しながらフラットな Folksonomy を構造化する方法を提案するのが本論文のねらいである.

3.　Folksonomy に関する仮説

この章では以降 Folksonomy を分析する上で基礎となる,Folksonomy の性質に関する我々の仮説を展開する. 以下は我々の仮説の概略である.

タグ間のドキュメントベースの共起率とユーザベースの共起率の双方を計算することでタグどうしの関係性の推定精度が大幅に向上する

まず, ドキュメントベースの共起率とユーザベースの共起率の定義について説明していく.

3.1　2 タグ間の共起率

3.1.1　ドキュメントベースの共起率

キーワードどうしの共起率は文書検索や文書分類などの分野で頻繁に利用される. 通常は 1 つのドキュメント内に 2 つのキーワードが同時に出現する確率を意味し, この値が高いほど 2 つのキーワードの関連性が高いとみなされる.

ここではタグ間のドキュメント共起率を ”1 つのドキュメントに 2 つのタグが同時にラベリングされる確率 ”と定義する. 共起率の計算方法としては以下に示す AEMI(Augmented Expected Mutual Information) を用いる [7].

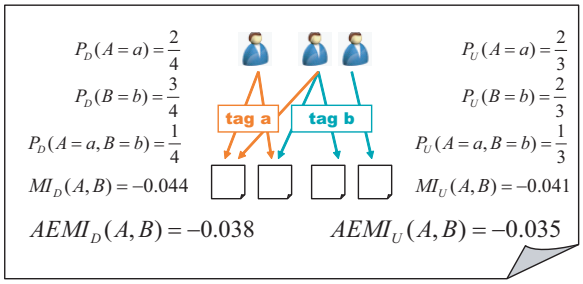


図 6　Co-Occurrence Rate between Tag a and b

$$MI_D = P_D(a,b)\log\frac{P_D(a,b)}{P_D(a)P_D(b)}$$
$$AEMI_D = MI_D(A=a,B=b)+MI_D(A=\bar{a},B=\bar{b}) - MI_D(A=a,B=\bar{b})-MI_D(A=\bar{a},B=b)$$

式中に出現する A,B は 2 つのタグ a,b に関連する変数である.$P_D(A=a)$ は任意のドキュメントにタグ a がラベリングされる確率を表す. 逆に $P_D(A=\bar{a})$ は任意のドキュメントにタグ a がラベリングされない確率を表す.$P_D(A=a,B=b)$ は任意のドキュメントにタグ a,b が同時にラベリングされる確率を表す.

MI_D はこれらの値から導き出したタグ a,b 間の共起率の 1 つの指標であり,$AEMI_D$ は MI_D を利用してより精細に共起率を算出している.

3.1.2　ユーザベースの共起率

ところで, 図 3 や図 6 で示されているように Folksonomy の構造の実体はユーザ, ドキュメント, タグの 3 種のノードによって構成される 3 部グラフである. したがってタグ間の共起率は前節のようなドキュメントノード上の共起率だけでなく, ユーザノード上の共起率という観点からも測れる.

以下, 前節と同様にユーザベースのタグ間共起率 $AEMI_U$ を定義する.

$$MI_U = P_U(a,b)\log\frac{P_U(a,b)}{P_U(a)P_U(b)}$$
$$AEMI_U = MI_U(A=a,B=b)+MI_U(A=\bar{a},B=\bar{b}) - MI_U(A=a,B=\bar{b})-MI_U(A=\bar{a},B=b)$$

$P_U(A=a)$ は任意のユーザがタグ a を用いたことがあるか否かの確率を表す.$P_D(A=a,B=b)$ は任意のユーザがタグ a とタグ b の両方を用いたことがあるか否かの確率を表す.

このように,Folksonomy が 3 部グラフ構造であることを利用してタグ間の共起率をドキュメントベースとユーザベースの双方から測定することができる (図 6). 我々はこの 2 つの異なる共起率に利用して以下のような仮説を展開する.

3.2　仮　　説

3.2.1　Synonym 関係

以下は Synonym 関係のタグに関する我々の仮説である.

2 つのタグが Synonym 関係の場合, それらのユーザベース共起率 ($AEMI_U$) はドキュメントベース共起率 ($AEMI_D$) に比べて大幅に低い傾向にある

2 章でも述べたが,business と biz のように同じ意味で別表記のタグを Synonym 関係にあるという.Folksonomy において Synonym タグが生じるのは, 異なるユーザ間でどのタグを使うかに関するコンセンサスがとれていないことに起因する. したがって Synonym 関係にある 2 タグが同一ユーザによって用いられている可能性は低い. それに対し,Synonym タグどうしは意味は同じなので同一のドキュメントにラベリングされている可能性が非常に高い. 一般にユーザベースの共起率とドキュメントベースの共起率は正の相関関係にあると考えられるので, この特徴を利用すれば Synonym 関係の判定を高精度で行うことができるだろう.

3.2.2　Conflict 関係

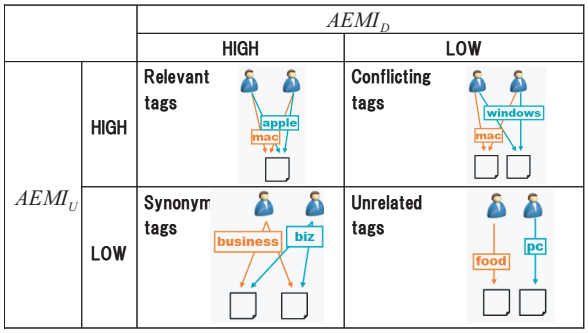
本論文中では意味的に対立もしくは競合している 2 タグを Conflict 関係にあると定義する. 例えば man と woman,mac と windows など同レベルの異なる事象を表している場合, これらは Conflict 関係にある.Synonym 関係同様,Conflict 関係に対しても 2 つの共起率をもとに次の仮説を立てた.

2 つのタグが Conflict 関係の場合, それらのユーザベース共起率 ($AEMI_U$) はドキュメントベース共起率 ($AEMI_D$) に比べて高い傾向にある

この仮説は mac と windows のような対照関係にあるタグは同一ユーザによって利用されている可能性が高いが, 同一ドキュメントにラベリングされている可能性は低いという仮定に基づいている. ただしこれらのタグが同一ドキュメントに現れる確率は低いといっても一定数見込まれるため,Synonym 仮説ほどには顕著な特徴が現れないのではないかというのが仮説を立てた段階での予想である.

3.2.3　Relevant 関係

最後に, 意味的な相関度が高い 2 タグのうち Synonym 関係でも Conflict 関係でもないものを Rele-



$AEMI_D$ indicates co-occurrence rate of two tags on the same document
$AEMI_U$ indicates co-occurrence rate of two tags on the same user

図 7　Our Assumptions about Folksonomy Tag Relations

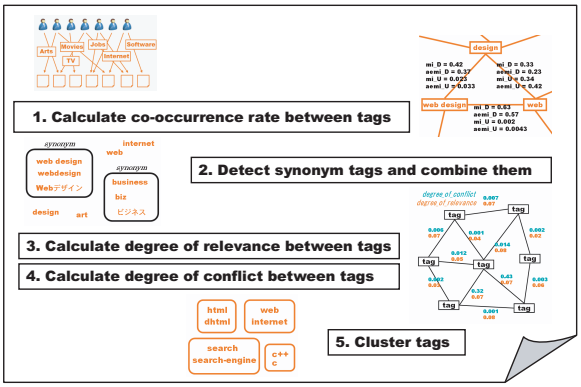


図 8　Our Tag Clustering Method

vant 関係と定義する.Relevant 関係の例としては software と programming,food と recipe などがあげられる.Relevant 関係に関する仮説は以下のとおりである.

2 つのタグが Relevant 関係の場合, それらのユーザベース共起率 ($AEMI_U$) とドキュメントベース共起率 ($AEMI_D$) は共に高い傾向にある

この仮説は意味的な相関度の高い 2 タグは同一ユーザにも利用されやすく, また同一ドキュメントにもラベリングされやすいという仮定に基づいている.

図 7 は Folksonomy のタグの関係性に関する我々の仮説を 2 つの共起率を軸にまとめたマトリックスである. これらの仮説の妥当性については 5 章にて評価実験を行う.

4. Folksonomy タグのクラスタリング

4.1 タグクラスタリングの必要性

タグは Folksonomy を扱う上で最も重要な存在だが,Taxonomy のカテゴリのような意味的な構造化が全くなされていない.del.icio.us を例にとると数万種類を越えるタグが用いられており, 中には特定のユーザしか用いていないような希少タグも存在するし,news や blog のように頻繁に用いられるタグも存在する. 先に述べた Synonym 関係のように意味が同一のタグや Relevant 関係のように意味的な相関度の高いタグどうしも区別なくフラットに入り混じっている.

Folksonomy を Web Recommender などに応用する場合, 例えばドキュメントに付与されたタグをもとにそのドキュメントをベクトル化するといった用法が想定される. その場合も 1 種類のタグを 1 つの次元にそのままマッピングしていたのでは非常に計算効率が悪いので,Synonym 関係のタグや Relevant 関係のタグをクラスタ化して次元数を圧縮する必要があるだろう.

この章では 3 章で述べた我々の仮説の応用例として,Folksonomy タグをクラスタリングする方法を示す.

4.2 タグクラスタリングの手順

図 8 はタグクラスタリングの手順を示している. 以下, これらの手順について順に説明していく.

4.2.1 共起率の計算

タグ間の共起率を求める. 共起率は前章で述べた $AEMI_D$ と $AEMI_U$ の 2 つについてそれぞれ求める. 計算の際はドキュメントベースの共起もユーザベースの共起も全く見られない 2 タグ間については計算しないなどして, 全く無関係だと思われる 2 タグ間の計算を省く.

4.2.2 Synonym タグの融合

3.2.1 で述べた Synonym タグに関する我々の仮説をもとに Synonym タグを検出する. 具体的には $AEMI_D$ が十分に高く $AEMI_U$ が十分に低くなるような 2 タグを Synonym 関係とみなすのだが, その判定を行うためには適切な閾値を設定してやる必要がある. 今回はあらかじめ人間が作った訓練データから決定木を生成する教師付き学習を用いて Synonym の判定を行う. 決定木生成の詳細と判定精度の評価は次章を参照のこと.

Synonym 関係と判定されたタグどうしは 1 つにまとめられ, 以降のフェーズでは同一タグとして扱われる.

4.2.3 相関度の計算

我々のクラスタリングではより意味的な相関度が高いタグどうしがクラスタ化されることをめざす. このフェーズでは 3.2.3 で述べた仮説をもとにタグ間の相関度 (degree of relevance) を計算する. 相関度を測る

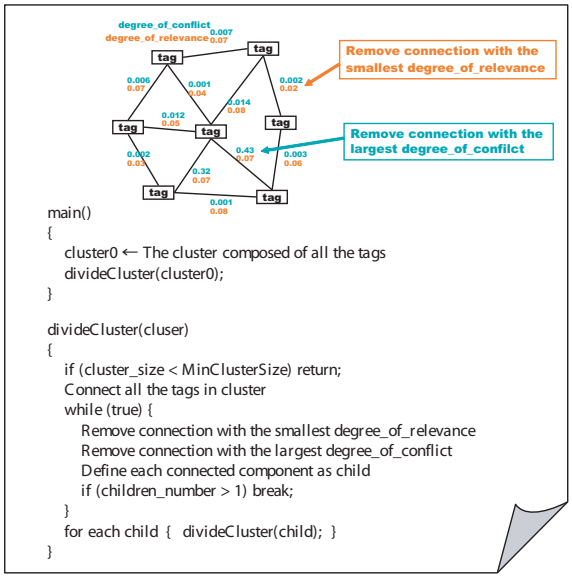


図 9　Our Tag Clustering Algorithm

基準値には以下の値を用いる.

$$relevance(A,B) = AEMI_D(A,B) \times AEMI_U(A,B)$$

4.2.4　対立度の計算

相関度とは逆に, 我々のクラスタリングでは意味的な対立度が高いタグどうしは別クラスタに分断されることをめざす. このフェーズでは 3.2.2 で述べた仮説をもとにタグ間の対立度 (degree of conflict) を計算する. 対立度を測る基準値には以下の値を用いる.

$$conflict(A,B) = \frac{AEMI_U(A,B)}{AEMI_D(A,B)}$$

4.2.5　クラスタリングの実行

タグ間の相関度と対立度をもとにクラスタリングを実行する. 図 9 はクラスタリングアルゴリズムを示す.

まず初めに全てのタグを 1 つの巨大なクラスタとみなし, これを再帰的に分割していく.Synonym タグは既にまとめられおり, このフェーズでは 1 つのタグとして扱う.1 つのクラスタを分割する手順は以下のとおりである.

まず, クラスタ内に含まれるタグ間の接続のうち最も相関度が低いものを取り除く. 次に最も対立度が高いものを取り除く. これをクラスタ内の連結成分が 2 つ以上に分解するまで続ける. さらに分割された子クラスタにも同様のアルゴリズムを再帰的に適用していく.

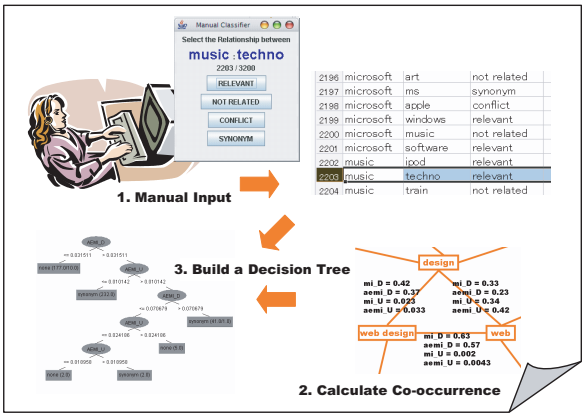


図 10　Evaluation Method

この分割アルゴリズムの特徴は同一クラスタ内における最小の相関度を高く, 最大の対立度を低く設定できる点にある. クラスタの分割はクラスタのサイズが閾値以下になった時点で止める.

5.　評価実験

この章では 3 章で述べた我々の仮説の検証を行う. 具体的にはドキュメントベースとユーザベースの 2 種類のタグ共起率を用いることで, 既存のドキュメントベースのみの場合に比べどれほど Synonym 関係などの判定精度が向上するかを検証する.

Folksonomy のサンプルデータは del.icio.us 上で 2006 年 7 月中に公開されているデータを用いた.21,000 人のユーザ,8400 種類のタグ,220,000 エントリ分のデータを使用した.

5.1　評価方法

図 10 は評価手順を示している. 以下, それぞれの手順について順に説明する.

5.1.1　人手によるタグ間の関係入力

この評価実験の目的は我々の仮説を用いることでタグ間の意味的関係の推定が高精度で行えるようになることを示すことである. したがって, 評価のためにはタグ間の意味的関係に対する正解をあらかじめ用意しておく必要がある. 今回我々は人間による手入力によってタグ間の意味的関係を定義した.

del.icio.us 上で用いられているタグのうち最も使用頻度の高い上位 200 個のタグ間の関係について Synonym,Relevant,Conflict,Not Related のうちから手動で選択し入力を行った. その際,$AEMI_D$ と $AEMI_U$

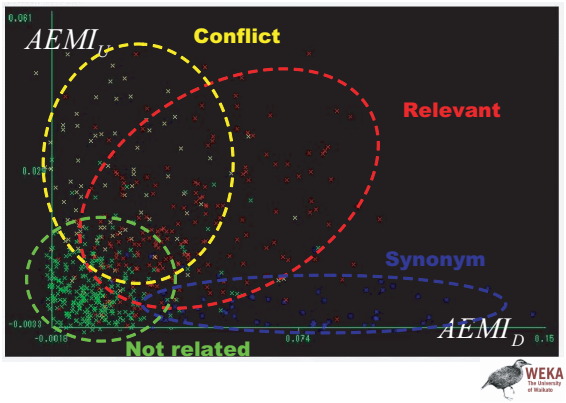


図 11　Distribution of the Training Set

Attributes \ Relations	Relevant	Synonym	Conflict	Not related	Total
$P_D(A), P_D(B), AEMI_D(A,B)$	13%	12%	16%	53%	35.0%
$P_U(A), P_U(B), AEMI_U(A,B)$	11%	1.3%	8.4%	43%	27.6%
$AEMI_D(A,B), AEMI_U(A,B)$	54%	72%	32%	65%	57.8%
$AEMI_D(A,B), AEMI_U(A,B), \frac{AEMI_D(A,B)}{AEMI_U(A,B)}$	66%	92%	52%	75%	70.8%

図 12　Classification Accurecy of Each Decision Tree

がともに十分に低い 2 タグについてははじめから Not Related とみなすことで入力の手間をおさえた.

入力値の偏りを防ぐため,4 人の人間が全く同じデータについて入力を行った.4 人のうち 3 人以上が同じ関係を入力した場合にのみ正解データに組み込んだ. そうでないデータに関しては慎重な議論の後再度入力を行った.

5.1.2　タグ間の共起率の計算

人力で入力された各タグの組み合わせに対し $AEMI_D$,$AEMI_U$ の 2 種類の共起率を計算し, 人力入力による関係データと合わせて訓練データとする.

5.1.3　決定木の生成

訓練データをもとに決定木を生成する. 今回はデータマイニングツールとして Weka3 [11] を用いた. 決定木の生成アルゴリズムには J48 を用いた.

5.2　評価結果

5.2.1　タグ間の関係の分布

図 11 は実験で用いた訓練データの統計的な分布である. 横軸と縦軸がそれぞれ計算によって求められた 2 タグ間の $AEMI_D$ と $AEMI_U$ を表し, 各点の色が人力で入力した 2 タグ間の関係性 (Synonym,Relevant,Conflict,Not Related) を表す.

Not Related は $AEMI_D$ と $AEMI_U$ が低い領域に集中し,Relevant は逆に右上方向に広く広がっているのがわかる.Conflict は Relevant の領域と大きく被るものの, 全体的に $AEMI_D$ が低い領域に分布している. 最も特徴がわかりやすいのが Synonym で,$AEMI_U$ が低い領域に独自の塊を作っている.

5.2.2　決定木の分類精度

我々は $AEMI_D$ と $AEMI_U$ の 2 種類の共起率の併用による分類精度の向上を示すため, 訓練データの入力属性を様々に変えて複数の決定木を生成し, それらの分類精度の比較を行った.

図 12 は各々の決定木の分類精度の比較表である. 例えば一番左上のマス (13% のマス) は $P_D(A)$,$P_D(B)$,$AEMI_D(A,B)$ の 3 つの属性をもとに生成した分類木が Relevant 関係の 2 タグを Relevant 関係だと正しく判定する確率が 13% であることを意味する.

これを見ると, ドキュメントベースの情報のみを使った決定木の総合分類精度が 35.0% , ユーザベースの情報のみを使った決定木の総合分類精度が 27.6% に対し $AEMI_D$ と $AEMI_U$ の双方を利用した場合の分類精度は 57.8% と極めて高くなっているのがわかる. とくに Synonym 関係の分類精度の向上が著しい.

また, 属性値に $AEMI_D$ と $AEMI_U$ の他に $\frac{AEMI_U(A,B)}{AEMI_D(A,B)}$ を加えて決定木を生成すると,Synonym 関係の分類精度と Conflict 関係の分類精度がさらに大きく向上する. これは Synonym 関係と Conflict 関係の判定には $AEMI_D$ と $AEMI_U$ の大小比が大きな意味を持つという我々の仮説を裏付けている.

6.　関連研究

Peter Mika は Folksonomy を Actor-Concept-Instance から成る 3 部グラフ構造と定義した [4].Peter Mika はこれを従来の Concept-Instace 型の 2 部グラフ構造のオントロジに社会的な次元 (Social Dimension) が加わったものだと主張し,Folksonomy をもとにオントロジを構築する可能性を示した.

丹羽らも同様に Folksonomy の 3 部グラフ構造に注目した [1]. 丹羽らは Folksonomy によって従来の User-Instance 型の 2 部グラフ構造からなる協調フィルタリングを拡張し, 高精度の Web 推薦システムを構築する手法を提案した.

Scott Golder らは Folksonomy(彼らは Collabo-

rative Tagging という言葉を使った) の構造を解析し,Folksonomy 特有の性質の発見を試みた [6]. 彼らは Folksonomy におけるユーザ行動, タグの使用率, 使われているタグの種類, ブックマークの人気などに関して興味深い法則を発見した. 本論文で提唱した Folksonomy タグの 2 種類の共起率に関する仮説は彼らの発見した法則をさらに発展させ, 補う形となっている.

大向らは Folksonomy におけるユーザどうしのつながり (Personal Network) に注目し,Personal Network を活用する SBS の新しい仕組みを提案した [3].

7. ま と め

我々は Folksonomy の 3 部グラフ構造から導かれる 2 種類のタグ共起率を併用することで,Synonym 関係などのタグどうしの意味的関係を高精度で推定する手法を提案した. この分類手法は Folksonomy の普遍的な性質に関する我々の仮説に基づいており, 評価実験を行って仮説の妥当性を検証した.

この分類手法を用いることで Folksonomy を扱う上で特有の問題である Synonym 関係の判定を自然言語処理を利用すること無く 90%以上の精度で行うことに成功した.

文　　献

[1] Satoshi Niwa, Takuo Doi, Shinichi Honiden : Web Page Recommender System based on Folksonomy Mining, *3th International Conference on Information Technology : New Generations*, (2006).

[2] Larry Page, Sergey Brin, R. Motwani, T. Winograd : The PageRank Citation Ranking: Bringing Order to the Web, (1998).

[3] Ikki Ohmukai, Masahiro Hamasaki, Hideaki Takeda : A Proposal of Community-based Folksonomy with RDF Metadata, *4th International Semantic Web Conference*, (2005).

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

[5] J Golbeck, B Parsia, J Hendler : Folksonomies-Cooperative Classification and Communication Through Shared Metadata, (2004).

[6] Scott Golder, Bernardo A. Huberman (HP Labs) : The Structure of Collaborative Tagging Systems, *Journal of Information Science, 32(2). 198-208*, (2006).

[7] Chan, P.K. : A non-invasive learning approach to building web user profiles, *KDD-99 Workshop on Web Usage Analysis and User Profiling*, (1999).

[8] digg : http://digg.com/

[9] del.icio.us : http://del.icio.us/

[10] Google : http://google.com/

[11] Weka3 : http://www.cs.waikato.ac.nz/ml/weka/