Vol.2015-MPS-102 No.5
2015/3/3

タグ情報における階層と非階層関係が共存したグラフ構造の抽出

呂ひろし[†1]　　鈴木泰博[†1,2]

ユーザが単語（タグ）をメタデータとして対象のアイテムに関連づけることのできるタクソノミーシステムは、大規模なデータを分類や整理をする手段として確立されてきた。フラットタクソノミーでは任意の単語がつかわれるため、タグ同士には明示的な関係性が存在しない。この論文ではフラットタクソノミーのタグを階層、および非階層関係を含むグラフ構造へ書き換える手法を提案し、そのアルゴリズムを説明したのち、実データに存在するこのグラフ構造のいくつかの特徴を示す。

Extracting a Graph Structure with Both Hierarchical and Nonhierarchical Relationships via Tagging Information

HIROSHI RO[†1]　YASUHIRO SUZUKI[†2]

Taxonomy systems - systems that allow users to annotate items with string formed metadata - have been established as a reasonable way to label and organize large collections of data for years. Due to the uncontrolled vocabulary, tags in a flat taxonomy system have no explicit relations and vary. In this paper, we introduce an algorithm for converting a set of tags in a flat taxonomy system into a hierarchical and nonhierarchical relationships coexisted in graph structure. We discuss the algorithm first, and then we show some features of the graph structure in real data

1. はじめに

本研究の目的はタグの情報から階層構造を抽出する手法を提案し、この手法に基づいて作られる構造の特徴を調べる事である。

タクソノミーとは、従来のフォルダー型の整理方法ではなく、対象のアイテム（ブックマークや論文など）について、「タグ」と呼ばれるユーザ選択による自由な文字列のメタデータを関連づけること（タグ付け）で整理することである。多数のユーザが個別に対象をタグ付けすることの出来るシステムをソーシャルタギングあるいはフォークソノミーと呼ぶ。これらは特に、急速に増大するウェブ上の情報を整理するための手法として、近年一般化した物である。フォークソノミーではタグ同士に明示的な関連がないため、タグデータ内の構造を調べる手法は以前から研究されてきた。タグ同士の共出現に基づく手法 [Kipp Campbell, 2006] や、頻繁に使われるタグ間の相似度を用いた手法 [Heymann Garcia-Molina, 2006]等が有名である。

これらの研究はマクロ的に多数のユーザ同士で共通して現れるパターンに焦点がおかれ、各ユーザの持つタグの構造をミクロ的に注目する物はほとんどない。本研究は、各ユーザのタグデータの構造を抽出する方法を提案し、このように得られる構造の特徴を調べた。

†1 名古屋大学

†2 慶応大学

2. 手法

T = {T1, …, Ti, …, Tn} と I = { I1, …, Ii , …, In } を Ti がタグ、
Ii がアイテムである集合とする。

I(Ti) = {Is, It, …} をタグ Ti にタグ付けされたアイテムを全て含む集合とし、T(Ii) = {Ts, Tt, …} がアイテム Ii が関連づけされたタグを全て含む集合とする。|I(Ti)| を Ti がタグ付けしたアイテムの総数とし、n(Ti) と書く。 |I(Ti) ∩ I(Tj)| を Ti とタグ Tj の両方にタグ付けされたアイテムの数とし、co(Ti , Tj) と書く。r(Ti , Tj) を式 (1) で定義し、タグ Ti と Tj の重なりの度合いとする。

$$r(Ti, Tj) = \frac{co(T_i, T_j)}{n(T_i)} \tag{1}$$

しきい値 δ を常に 1.0 と定める。A を式 (2) で定義される二項関係とし、A を降順の階層関係と呼ぶ。もし (Ti, Tj) が A に含まれるなら、Ti を Tj の先祖、Tj を Ti の子孫と呼び、Ti > Tj あるいは Tj < Ti. と書く。D(Ti) を Ti の全ての子孫を含む集合とし、 A(Tj) を Tj の全ての先祖を含む集合とする。

$$A = \{(T_i, T_j) \mid r(T_i, T_j) \geq \delta \wedge \delta > r(T_j, T_i) \wedge co(T_i, T_j) > 0\} \tag{2}$$

P を式 (2) で定義される A の部分集合とし、P を隣接の降順階層関係、あるいは単に階層関係と呼ぶ。もし (Ti, Tj) が P に含まれるならば、Ti を Tj の親、Tj を Ti の子と呼び、Ti / Tj あるいは Tj \ Ti と書く。

$$P = \{(T_i, T_j) | (T_i, T_j) \in A \wedge D(T_i) \wedge A(T_j) \neq \emptyset\} \quad (2)$$

F を式 (3) で定義される二項関係とし、F 非階層あるいは並列の関係と呼ぶ。もし (Ti, Tj) が F に含まれるなら、Ti を Tj の友達と呼び、Ti – Tj と書く.

$$F = \{(T_i, T_j) \mid \delta > r(T_i, T_j) \wedge \delta > r(T_j, T_i) \wedge co(T_i, T_j) > 0\} \quad (3)$$

G = (T, E) を、E = P∪F とするグラフとし、P を有向辺と呼び、F を無向辺と呼ぶ。このように作られるグラフは、マルチルートや三頂点閉路を持たないなどの特徴がある。このグラフを、有向辺を矢印で書き、無向辺を直線で書くと、Table 1 のタグデータを Figure 1 の構造に書き換える事が出来る。さらに、階層構造の度合いを示すために値 h を式 (4) として定義した。

$$h = \frac{|P|}{|E|} \quad (4)$$

Table 1: タグデータの一例

T_i	$I(T_i)$
T1	I1, I2, I3, I4, I5, I7, I8, I9
T2	I2, I5
T3	I3, I4, I5, I8, I9
T4	I4, I5, I9
T5	I3
T6	I6, I7, I8, I9
T7	I6, I7

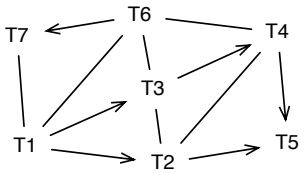

Figure 1 Table 1 のデータから作られる構造

3. 実験

この研究では、提案した手法によって作られる構造の特徴を調べるために主に delicious.com の実データを用いて、二つの実験を行った。最初の実験は各ユーザの |T|, |E|, |P|, |F| そして h の値の分布を調べ、Table 2 にその一覧をまとめた。その次に、実データに存在する構造の多様性を調べるため、ユーザ間での特定のタグのペアの構造の違いを調べた。

本実験に用いられたデータセットは delicous.com でランダムに得られた 17721 ユーザ分のタグ情報である。本データセットは 2014 年の 11 月から 12 月にかけて入手した。その際 |E| > 1 のユーザのみが選ばれ、各ユーザが持つリンクの数は最小 2 から最大 108602 であった。

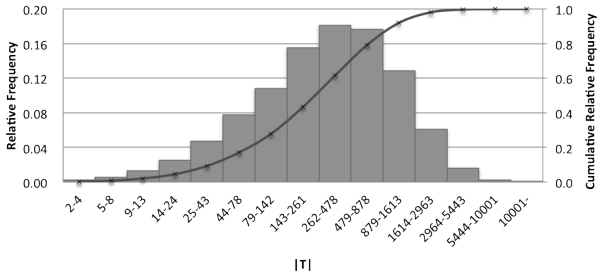

Figure 2: 値|T|の片対数度数分布図。平均値が 486.9、中央値が 237、標準偏差が対数スケールで 1.95、±2σ が 95.68%を占める。

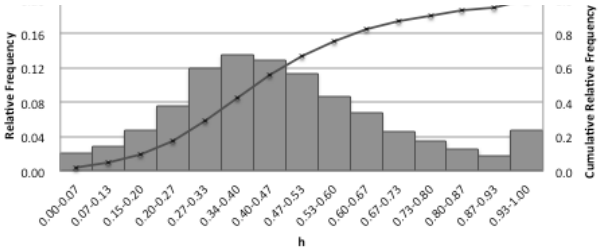

Figure 3: 値 h の度数分布図。平均値が 0.46、中央値が 0.44、標準偏差が 0.22、±2σ が 93.67%を占める。h = 1.0 が 4%を占める。

3.1 分布パターン

最初の実験では、まずユーザ間の |T| の分布を調べた。そのうち |T| の最小値は 2 であり、最大値は 16852 であった。平均値は 595.97 であり、中央値は 328 であった。これらの数値の特徴から、|T| の対数スケールでの分布を調べた。|T| の底を 2 とした対数スケールでは、最小値は 1 であり、最大値は 14.0 であった。平均値は 8.15、中央値は 8.36、標準偏差は 1.95、±2σ は 95.47%であった。この分布

Table 2: 分布に関する値の一覧。
h 以外は対数スケールである。

	|T|	|E|	|P|	|F|	h
Min	1.00	1.00	0.00	0.00	0.00
Max	14.04	19.40	17.01	19.31	1.00
Mean	8.15	9.08	7.83	8.10	0.46
Median	8.36	9.45	8.14	8.58	0.44
σ	1.95	3.19	2.85	3.56	0.22
μ-2σ	42.18%	41.72%	41.84%	40.43%	54.51%
μ-1σ	29.45%	28.95%	29.12%	27.26%	41.50%
μ+1σ	38.65%	39.39%	38.52%	40.25%	28.86%
μ+2σ	53.28%	53.76%	53.37%	55.04%	39.16%
±1σ	68.09%	68.34%	67.64%	67.51%	70.36%
±2σ	95.68%	95.49%	95.21%	95.47%	93.67%
Skewness	-0.50	-0.48	-0.41	-0.47	0.52
Kurtosis	0.12	-0.03	-0.19	-0.32	-0.02

の歪度は-0.5 であり、尖度は 0.12 であった。これらの特徴から、|T|の分布は対数正規分布であると考えられる。この分布のヒストグラムは Figure 2 で示した通りである。次に |E|, |P|, |F| の分布を調べたが、それぞれの分布の特徴は |T| と同様であった。この事は、タグの数を決定するユーザの行動パターンは、ユーザの持つグラフ構造の枝の数をも決定するためであると考えられる。

続いて、h の値の分布を調べた。その分布のヒストグラムは Figure 3 で示した通りである。|T| の片側対数の分布と比べると全体が左側に傾いており、また h = 1.0 付近には二つ目の峰がある。この事から、大多数のユーザは階層と並列の関係の混ざった構造を一般的に使うが、少数のグループのユーザは階層構造のみのタグ構造を意図的に使用する事がわかる。

値 h の特徴をより詳しく調べるために、|T|と h の二次元ヒストグラムを Figure 4 に作成した。このグラフには二つの特徴があることが分かる。最初の特徴は、度数分布の形が全体的に傾いていることである。この傾きは、タグの数が増えると、タグのグラフ構造内の階層の割合が減る事を示している。これは、タグの数がより増えると、階層構造を保つために常に特定のタグを一緒に同じリンクへのタグ付が減る事をあらわしている。これはユーザが意図的に作った傾向であるかはまだ断定出来ないが、もしこの傾向がユーザによって意図的に作られたものでないとすると、フォークソノミーのシステム自体が特定のタグが常に同時に使われるように機能的にサポートしていないことに起因していると考える事も出来る。二つ目の特徴は、|T| ではおよそ 14 から 52、h ではおよそ 0.9 以上の空間に一つ小さなピークがある事である。このピークは、h のヒストグラムで見られた物と同じ物であり、ほぼ階層構造のみの構造を意図的に利用するユーザのグループを示す。この小さなピークの占める空間から、このユーザグループは主に 15 から 50 個程度のタグしか持っていない事がわかる。この数は全体の |T| の数の分布からすれば比較的小さな数であり、この数がそれほど大きくない理由としては、第一の特徴で取り上げたフォークソノミーのシステムが階層構造を作るタグづけの行動パターンを特にサポートしていない事と関連している可能性も考えられる。

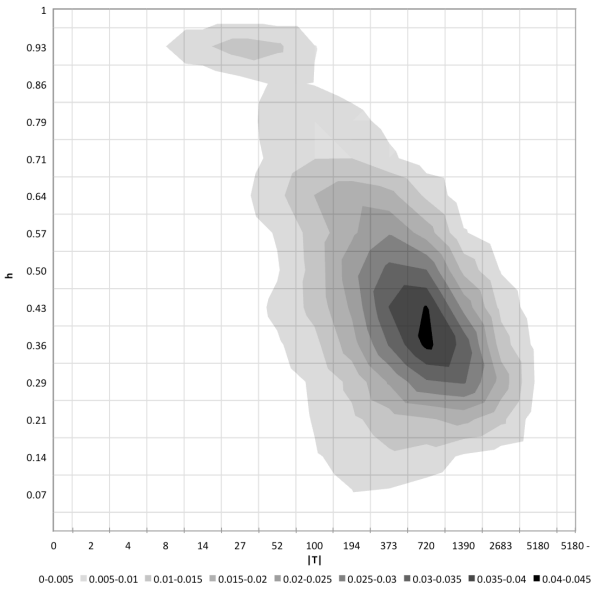

Figure 4: |T|と h の二次元度数分布。X 軸は Figure 2 の X 軸と同様に|T|の対数スケールであり、Y 軸は Figure 3 の Y 軸と同様に h の値である。

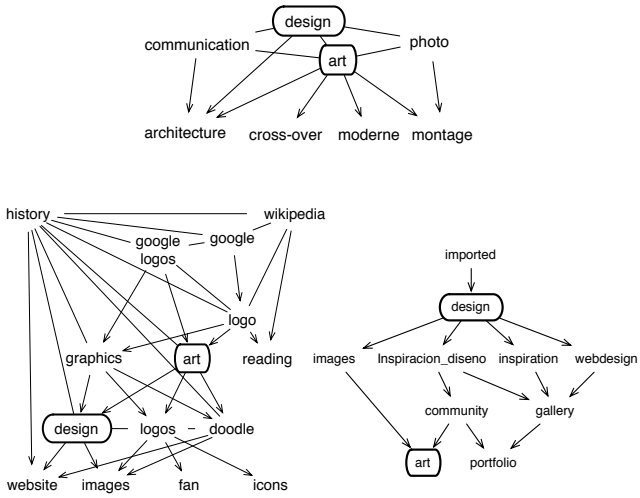

Figure 5: 'art'と'design'を含む構造の実例

3.2 構造パターン

二つ目の実験として、データセット内に存在する実際のタグのパターンを調べた。まずはタグ'art' と 'design'を例として取り上げた。全ユーザのうち、9385 ユーザが'art'のタグを、13004 ユーザが'design'のタグを使い、5749 ユーザがその二つのタグを一度以上同時に使った。この 5749 のユーザのうち、65 ユーザ（1.13%）が 'art' >'design'として、520 ユーザ（9.05%）が 'art'<'design' として、5164 ユーザ（89.82%）が 'art'-'design' として使った。データセット内で実際に存在する 'art'と'design' 構造の例を Figure 5 に示した。

ユーザ間のタグの構造の違いをみるために、20 個のタグのペアのグループを三つ選んだ。その内訳は、1) 最も多くのユーザに使われたタグ 'design' （13005 ユーザ）と共に最も良く使われた 20 個のタグのペア。2) 最も多くのユーザに共通して使われた 20 個のタグのペア（'design'を含む物を除く）。3) ともっとも多くのユーザに共通して使われた 20 個の階層関係のペア。それぞれのペアの構造のパターンの頻度を調べ、特に特徴的なパターンを示すペアを Table 3 に示した。

Table 3: 最も頻繁に使われたタグのペアのリストの一部。L|R とは L>R, L-R, L<R の合計である。

L	R	L|R	L>R	L/R	L=R	L\R	L<R	L>R(%)	L/R(%)	L=R(%)	L\R(%)	L<R(%)
design	webdesign	7212	491	444	6611	105	110	6.81%	6.16%	91.67%	1.46%	1.53%
design	inspiration	5804	669	601	5089	45	46	11.53%	10.35%	87.68%	0.78%	0.79%
design	art	5749	520	446	5164	57	65	9.05%	7.76%	89.82%	0.99%	1.13%
design	blog	5427	174	140	5219	29	34	3.21%	2.58%	96.17%	0.53%	0.63%
design	graphics	5161	722	615	4418	20	21	13.99%	11.92%	85.60%	0.39%	0.41%
design	typography	4614	1080	939	3528	5	6	23.41%	20.35%	76.46%	0.11%	0.13%
design	fonts	3919	744	573	3167	8	8	18.98%	14.62%	80.81%	0.20%	0.20%
design	architecture	3875	630	565	3226	19	19	16.26%	14.58%	83.25%	0.49%	0.49%
software	tools	6541	96	84	6167	246	278	1.47%	1.28%	94.28%	3.76%	4.25%
css	webdesign	6514	185	172	5586	671	743	2.84%	2.64%	85.75%	10.30%	11.41%
javascript	jquery	5605	1081	1000	4452	67	72	19.29%	17.84%	79.43%	1.20%	1.28%
ajax	javascript	5150	62	61	4343	622	745	1.20%	1.18%	84.33%	12.08%	14.47%
javascript	js	2610	1253	1146	1341	15	16	48.01%	43.91%	51.38%	0.57%	0.61%
music	mp3	3983	1229	1128	2700	51	54	30.86%	28.32%	67.79%	1.28%	1.36%
video	youtube	4253	1217	1136	3011	25	25	28.62%	26.71%	70.80%	0.59%	0.59%
design	logo	2726	1198	1055	1525	3	3	43.95%	38.70%	55.94%	0.11%	0.11%
design	color	3728	1104	1006	2616	8	8	29.61%	26.98%	70.17%	0.21%	0.21%
mac	osx	3646	1102	1057	2403	140	141	30.22%	28.99%	65.91%	3.84%	3.87%
search	searchengine	2220	1097	1042	1099	24	24	49.41%	46.94%	49.50%	1.08%	1.08%
flash	actionscript	2096	1084	1030	1001	11	11	51.72%	49.14%	47.76%	0.52%	0.52%
linux	ubuntu	3919	1084	1033	2786	46	49	27.66%	26.36%	71.09%	1.17%	1.25%
design	typography	4614	1080	939	3528	5	6	23.41%	20.35%	76.46%	0.11%	0.13%
software	freeware	3686	1025	952	2652	9	9	27.81%	25.83%	71.95%	0.24%	0.24%
security	password	2137	1023	942	1098	15	16	47.87%	44.08%	51.38%	0.70%	0.75%
travel	flights	1359	1010	845	344	5	5	74.32%	62.18%	25.31%	0.37%	0.37%

以上の 60 組のタグのペアには三つの特徴的なパターンがあった。最初の特徴的な分布とは、並列関係が主に使われるが、片側の階層がもう片側の階層よりも圧倒的に多いことである。その例で最も顕著のペアは ‘design’ と’typography’ などである。約 77%のユーザが ’design’ – ‘typography’としてこのペアを使うが、約 25%のユーザが ’design’ > ‘typography’として使った。この分布を持つものの多くは大きいがテゴリーと小さいがテゴリーを示す物の組み合わせと考えられる。次に、並列関係と片方の階層関係がほぼ同じパターンだということである。’search’ と’searchengine’を例にあげると、共に 49.5%のユーザが’search’>’searchengine’または’search’-’searchengine’ としてこのペアを使った。最後に、一つ変わったパターンの例として、’travel’と’flights’ がある。75%近くのユーザが "travel">"flights"を使ったが、"travel"-"flights"を使ったのはわずか約 25%のユーザであった。

4. 結論

この研究は、フォークソノミー内の各ユーザの持つタグの情報をグラフ構造として書き換える手法を提案した。この手法を用いる事で各ユーザの持つタグの階層を見る事が出来、その構造に対して分布とユーザごとの構造の違いを調べた。分布では、より多くのタグを持つユーザは、より少ない階層の割合を持つ傾向が一般的である事がわかった。また、少ないタグを持ち、ほぼ階層構造のみを用いるユーザは少数であるが存在する事がわかった。ユーザの持つ特

徴として、同じ二つのタグのペアをユーザによっては全く異なる構造のもとで使うことがあることがわかった。

5. おわりに

本研究で提案した手法によって、フォークソノミー内のユーザの様々な構造を局地的に取り出す事が可能であった。階層と並列の共存したグラフ構造によってタグデータを分析する事は、この研究で調べた内容以外にも様々な可能性があると考えられる。この手法のより一般的に展開するとともに、より細かくユーザの持つ構造パターンを解析することがタグデータのより良い理解と応用につながると考えられる。

参考文献

1) BegelmanG, KellerP, SmadjaF. (2006). Automated tag clustering: Improving search and exploration in the tag space. Collaborative Web Tagging Workshop at WWW2006 (ページ: 15-33). Scotland,: Edinburgh,.
2) GolderAScott,, HubermanA.Bernardo,. (2006). Usage patterns of collaborative tagging systems. Journal of information science, 32 (2), 198-208.
3) HeymannPaul, Garcia-MolinaHector. (2006). Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical Report. Stanford.
4) KippM, CampbellD.G. (2006). patterns and inconsistencies in collaborativve tagging systems: an examination of tagging practices,. american society for information science and technology.
5) LaniadoDavid, EynardDavide, ColombettiMarco. (2007). Using WordNet to turn a folksonomy into a hierarchy of concepts. Semantic web application and perspectives-fourth italian semantic web workshop, (ページ: 192-201).
6) SchmitzPatrick. (2006). Inducing ontology from flickr tags. Collaborative Web Tagging Workshop at WWW2006 (ページ: 50). Scotland: Edinburgh.
7) SchoefeggerK, TammetT, & GranitzerM. (2013). A survey on socio-semantic information retrieval. Computer Science Review, 25-46.
8) SmithGene. (2005年11月8日). Tagging tags to make synonyms. 参照日: 2014 年 12 月 7 日, 参照先: atomiq.org: http://atomiq.org/
9) YooChoi, K., Suh, Y., & Kim, G.D.,. (2013). Building and evaluating a collaboratively built structured folksonomy. Journal of Information Science, 39 (5), 593-607.