概念の上下関係に基づく階層構造を用いたラベルの集約

百武靖悟[1,a)]　古川哲也[2]

概要：近年オンライン上のデータの分析が広く行われるようになった。このようなデータの分析は流行の検出に欠かせないものであり、トレンド機能や企業の宣伝など様々な場面で使われている。本研究は、データの発生件数の増減と概念の上下関係を表す階層構造に基づいて、流行の検出を行う。ラベルが付されたデータを用い、データ数が増加している複数のラベルを最小個数のラベルで集約する。集約によりデータが増加している複数のラベルを 1 つの流行としてまとめることができる。

Detecting Summarizing Labels by Using a Conceptual Hierarchy

Seigo Hyakutake[1,a)]　Tetsuya Furukawa[2]

Abstract: Recently, online data analysis has spread rapidly, which is important for detecting trends. Online data analysis is used in various situations such as trend functions and corporate advertisements. This paper discusses detecting trend based on the conceptual hierarchy of labels. The labels of increasing data are summarized in the minimum number of labels to detect trend abstractly.

1. はじめに

近年、SNS の急激な広まりにつれて、SNS などのオンライン上のデータについての研究が広く行われるようになってきている。オンライン上のデータの解析はリアルタイムの人々の思考を探るための有用な手段として注目されている。SNS などのオンライン上のデータを用いて流行を検出する研究は年々増えてきており、各単語の出現頻度を探るものや、単語の関連性に着目したもの、インフルエンサによる影響力を分析するものといったものがある。たとえば、アメリカなどの選挙において、Twitter などの SNS を用いた戦略が使われる [1] など様々な方面でオンライン上のデータが用いられるようになってきた。このようにテキストに基づくものや、ユーザに基づくもの [2] など流行の検出は重要な役割を担っている。

流行の検出について、ある特定の単語の出現頻度を探り、今後の流行の予測を立てるもの [3] がある。ある特定の属性の言葉に対して、周期性を考えることで同様な出現頻度を持つものの予測を行う。周期性に注目することで長期的な解析においても予測を可能としている。言葉の関連性に着目した研究では、2 つの単語の出現頻度の因果関係を検討するもの [4] などがある。たとえば、SNS において、投稿内容の単語やハッシュタグに基づいた分析や Twitter のリツイートなどを用いた拡散機能に基づいた分析 [5][6] が行われている。

しかし、これらの研究はある特定のイベント前後での影響や、1 つの属性に関して限定的に絞られた単語に焦点を当てたものが多い。社会の流行を考える際、流行をとらえるには単語の単体精査だけでは不十分である。

単語を 1 つ 1 つ数えた場合、その単語自身の出現頻度の変化についてはわかるが、大きく流行をとらえることができない。例えば、「100m 走」や「200m 走」のような単語の出現頻度が増加しているとき、それぞれの出現頻度の増加の原因を個別で考えると大きく流行をとらえることはできない。データ数が増加していなくとも、「短距離」や「陸上競技」といったデータ数が増加している単語の上位概念でまとめることが適当である。このような場合、複数の単

[1] 九州大学大学院経済学府経済工学専攻
Department of Economic Engineering, Kyushu University, Fukuoka, 819-0395, Japan
[2] 九州大学大学院経済学研究院経済工学部門
Department of Economic Engineering, Kyushu University, Kyushu University, Fukuoka, 819-0395, Japan
[a)] 3EC19001G@s.kyushu-u.ac.jp

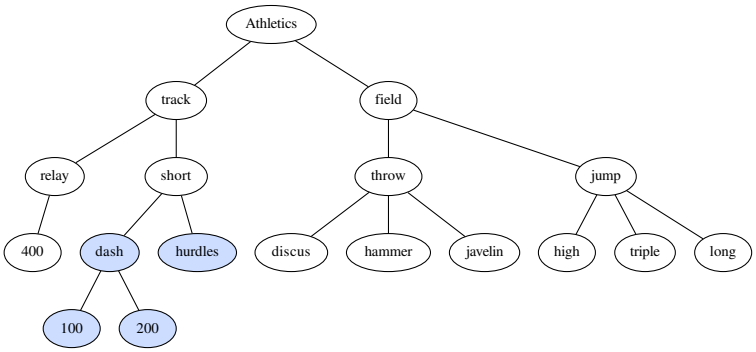

図 1 最小被覆ラベル

語を上位概念でまとめることで、その上位概念のデータ数が増加していなくても流行していることをとらえることができる。このように粒度の異なる概念をまとめることで大きく流行をとらえることができる。

本研究では、内容や意味を表すラベルが付されたデータを対象とする。複数のラベルでデータ数が変化していたとき、大きく流行をとらえるために複数のラベルを集約する。複数のラベルをまとめるとき、内容の類似度や 2 つのラベルの関係性に基づく手法があるが、本稿ではラベルの粒度に基づいてラベルを集約する。そこで、概念の上下関係に基づく階層構造を用いることで複数のラベルの集約を行う。

本稿の構成は以下の通りである。2 節では集約におけるラベルの関係性について述べる。3 節では大きく流行をとらえためにラベル集合を集約するラベルについて議論する。4 節では 1 つの階層内に複数の流行があり階層を分割する際の分割の性質について検討する。5 節では階層を分割するアルゴリズムと、分割の結果が唯一であることを示す。6 節では本稿のまとめとモデルの拡張について述べる。

2. ラベルの集約

SNS の投稿などにおける 1 件のデータをオブジェクト o とし、o にはその内容によってラベルが付されているものとする。オブジェクト o に付されているラベル集合を $L(o)$、ラベル l が付されているオブジェクト集合を $\bar{l} = \{o \mid l \in L(o)\}$ とする。ラベルには属性ごとに概念の上下関係を表す階層が与えられているものとする。本稿では議論を簡単にするために属性は 1 つであると仮定する。ラベル l がラベル l' よりも上位もしくは同等であることを $l' \preceq l$ で表し、上位である場合は $l' \prec l$ で表す。ラベル l がラベル集合 L 中のどのラベルよりも上位にあるときは、ラベル l はラベル集合 L の上位であるといい $L \prec l$ と表し、L の最上位が l であるときは $L \preceq l$ と表す。

本研究の目的は、ある二期間において、各期間で発生したデータ数の変化に基づきラベルをまとめることで、何が流行したかを特定することである。データ数の変化については二期間の同じラベルが付されているデータ数で比較をする。二期間 $i, j (i < j)$ におけるデータ数を $\bar{l^i}, \bar{l^j}$ とする。ある正の値 $\alpha(> 1), \beta(< 1)$ を用いて、$\beta|\bar{l^i}| \leq |\bar{l^j}| \leq \alpha|\bar{l^i}|$ であれば L のオブジェクト数の有意な変化はないとする。$\alpha|\bar{l^i}| < |\bar{l^j}|$ であればデータ数が増加、$\beta|\bar{l^i}| > |\bar{l^j}|$ であればデータ数が減少している。

複数のラベルのデータ数が増加しているとき、データ数の増加を各々で検討すると、そのラベル自身のデータ数の増加していないが概念として流行している場合に大きく流行をとらえるができない。このとき、データ数が増加しているラベルをまとめることで大きく流行をとらえることができる。データ数が増加しているラベルが複数存在したとき、そのラベルをまとめることで流行を検出する。このようにラベルでまとめることをラベルを集約するという。

例 1 図 1 は陸上競技について概念の上下関係を表す階層構造である。データ数が増加しているラベル集合が $L_1 = \{100, 200, dash, hurdles\}$ であるとき、これらのラベルのデータの増加を個別で考えると、それぞれの単語の出現頻度の変化はわかるが大きく流行をとらえることができない。これらのラベルを 1 つにまとめると、短距離に関する項目の各データ数が増加しているため、$short$ というラベルのデータ数が増加していなくても、短距離の項目が流行しているととらえることができる。よって、L_1 を $short$ というラベルで集約することで、大きく流行をとらることができる。

このようにラベルを集約する際は、上位のラベルによって集約される。

3. ラベル集合を集約するラベル

本節では、ラベル集合が与えられたとき、複数のラベルを集約するラベルがどのラベルであるか検討する。ラベル集合を集約するラベルを検出することで、大きく流行をとらえることが可能となる。

例 1 で示したように下位の概念を集約するラベルを集約ラベルとする。集約ラベルは複数のラベルをまとめるため、複数のラベルの共通な祖先であるラベルが集約ラベル

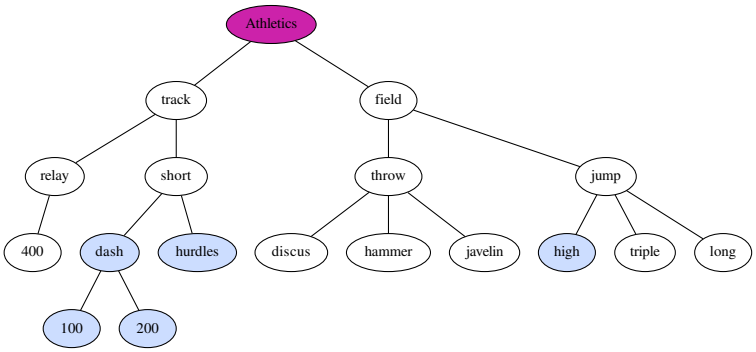

図 2 最小被覆ラベルで集約できない例

の候補となる。

定義 1 ラベル l は、ラベル集合 L の上位であるとき、L の被覆ラベルという。$l' \prec l$ となる L の被覆ラベル l' が存在しなければ、被覆ラベル l は L の最小被覆ラベルという。

最小被覆ラベルは共通な祖先のうち最も下位にあるラベルである。このラベルは集約ラベルの候補となる。

例 2 $L_1 = \{100, 200, dash, hurdles\}$ の被覆ラベルは $Athletics, track, short$ である (図 1)。 このとき、L_1 の最小被覆ラベルは $dash$ となる。

ラベル l がラベル集合 L 中のラベル l' の祖先であり、L 中のラベル l'' が $l'' \prec l'$ であるならば、l は l'' の祖先でもある。すなわち、l が l'' の祖先であるかどうかは、l が l' の祖先であるかどうかによって判断することができる。すなわち、l' 以下のラベルの被覆は、l' の被覆で考えることができる。

定義 2 ラベル集合 L 中のラベル l_r は上位となるラベルが存在しないとき L の代表ラベルであるといい、L のすべての代表ラベルの集合 L_r は L の代表ラベル集合であるという。ラベル $l \in L$ は $l \preceq l_r$ となる L の代表ラベル l_r で代表されるという。

ラベル集合 L 中の各ラベルはいずれかの代表ラベルで代表され、ラベル集合 L の被覆ラベルは代表ラベル集合 L_r の被覆ラベルである。よって、L の最小被覆ラベルは L_r の最小被覆ラベルであり、L を集約することは、L_r を集約することと等価である。

L_r の最小被覆ラベルは L の集約ラベルの候補となる。L が代表ラベルを 1 つしか持たないならば、その代表ラベルが L の集約ラベルとなる。2 つ以上の代表ラベルがあるとき、L_r は 1 つのラベルで集約すべきでない場合がある。代表ラベルが複数存在する場合は、ラベルが点在しているときに発生する。

例 3 $L_2 = \{100, 200, dash, hurdles, high\}$ であるとき、L_2 の最小被覆ラベルは $Athletics$ となる (図 2)。しかし、$Athletics$ で集約すると、L_2 中のラベル $high$ と $100, 200, dash, hurdles$ が離れすぎている。L_2 は $Athletics$ が流行しているとするのではなく、流行が 2 つあるととらえるべきである。

このように単純に共通な祖先で集約すると、複数の流行としてとらえるべきであるものを、1 つの流行としてとらえてしまう。よって、1 つの流行の範囲を探るために、ラベル同士がどれくらい離れているかを考えなければならない。ラベル同士が離れていると、各ラベルから最小被覆ラベルまでの距離が大きくなる。ラベル間の距離を最小被覆ラベルまでの距離で判断する。

本稿では、階層構造で親子関係にあるラベル間の距離はあらかじめ与えられているものとする。ラベル $l, l'(l < l')$ の距離は、l と l' 間の各ラベル間の距離の合計とする。ラベル間の距離は概念の距離を表している。

代表ラベル集合 L_r が 1 つのラベルで集約されるかどうかを、L_r 中のラベルと L_r の最小被覆ラベルの距離で判断する。距離が十分に小さければ、L_r 中のラベルは近いと判断できる。

定義 3 L の最小被覆ラベルと代表ラベルの距離が閾値 ρ 以下であるとき、ラベル集合 L は ρ で制限されるという。

ラベル集合 L が閾値 ρ で制限されていれば、L はまとまっていると判断する。

定義 4 ラベル集合 L の最小被覆ラベルは L の代表ラベルとの距離が閾値 ρ で制限されているとき、L の集約ラベルである。

例 4 親子のラベル間の距離を 1 であるとし、ρ が 1 であるとする。L_1 は最小被覆ラベルが $short$ であり、代表ラベル $dash, hurdles$ は ρ で制限されているので、L_1 の集約ラベルは $short$ である。L_2 は最小被覆ラベル $Athletics$ で集約すると、代表ラベル $dash, hurdles, high$ の距離が閾値より大きいため ρ で制限されておらず、$\{100, 200, dash\}$ と $\{high\}$ に分けるとそれぞれの最小被覆ラベル $short$,$high$

でまとめると ρ で制限されている。よって、L_2 の集約ラベルは $short, high$ の 2 つである。

4. ラベル集合の分割

流行が複数であるかどうかの判断を閾値 ρ で行う。ρ で制限されていないとき、複数の流行があるため、ラベル集合を 1 つのラベルでまとめることができない。このとき、流行を検出するためには、ラベル集合を 1 つの流行としてまとめられる大きさに分割しなければならない。本節ではラベル集合の分割についての性質を議論する。

分割の際に、ラベル集合 L の分割は代表ラベル集合 L_r の分割となるため、代表ラベル l_r 以下のラベルとなるラベル l は l_r と同じ分割に属する。よって、L の分割は代表ラベルの分割となっており、分割は垂直方向に下位の概念を伴っているため、分割は垂直に行われる。

定義 5 L の分割 $L_1, L_2, ..., L_m (L_i \cap L_j = \emptyset\ (i \neq j), \sum L_i = L)$ は、異なるラベル集合のラベル間に上下関係がないとき、L の垂直分割であるといい、L' を要素とする L の垂直分割が存在するとき、L' を L の垂直要素という。

例 5 $L_2 = \{100, 200, dash, hurdles, high\}$ の代表ラベル集合は $\{dash, hurdles, high\}$ である (図 2)。このとき代表ラベル集合はラベル集合 L_2 のラベルをすべて被覆している。よって、ラベル集合 L の分割は代表ラベル集合の分割である。また、垂直要素は $\{100, 200, dash\}, \{hurdles\}, \{high\}, \{100, 200, dash, hurdles\}$ である。

補題 1 ラベル集合 L の垂直要素 L_1, L_2 の和集合 $L_1 \cup L_2$ の最小被覆ラベルに被覆されるすべての L の垂直要素の和集合は、L の垂直要素である。
証明 L_2 は L の垂直要素なので、$L - L_2$ 中のラベルは L_2 中のラベルと上下関係がない。$L_1 \subset L$ なので L_2 中のラベルは $L_1 - L_2$ 中のラベルと上下関係がない。$L_2 \subseteq L_1$ のとき、L_2 を要素とする L_1 の垂直分割が存在するので、L_2 は L_1 の垂直要素であり、$L_1 \subseteq L_2$ の時も同様である。L_1, L_2 に包含関係がないとき、$L_1 \cup L_2$ 中の最小被覆ラベルに被覆されるすべての L の垂直要素の和集合を L' とする。$L - L'$ 中のラベルは最小被覆ラベルの下位ではないので L' のラベルと上下関係がない。$L' \subseteq L$ であるため、L' を被覆するラベルを代表ラベルとする垂直分割が存在するので、L' は L の垂直要素である。 (証明終わり)

例 6 $L_2 = \{100, 200, dash, hurdles, high\}$ の垂直要素 $\{100, 200, dash\}, \{hurdles\}$ の和集合 $\{100, 200, dash, hurdles\}$ は L の垂直要素である。また、垂直要素 $\{hurdles\}, \{high\}$ の和集合は $\{hurudles, high\}$ であり、和集合の最小被覆ラベルは $Athletics$ である。垂直要素 $\{100, 200, dash\}$ も和集合の最小被覆ラベル $Athletics$ に被覆されている。和集合の最小被覆ラベルに被覆される垂直要素の和集合 $\{100, 200, dash, hurdles, high\}$ は L の垂直要素である。

垂直分割を行う際に、距離が近いラベル同士の共通の祖先となっているラベル以外のラベルが集約ラベルの候補となることはない。このため、集約ラベルの候補となる代表ラベルが新たに発生することはない。

補題 2 ラベル集合 L, L の垂直要素 L' とその代表ラベル集合 L_r, L'_r に対し、$L'_r \subseteq L_r$ である。
証明 L' は垂直要素なので、L' の上位となるラベルは $L - L'$ 中に存在しない。L' の代表ラベルの上位となるラベルは $L - L'$ 中にないので、L' の代表ラベルは L の代表ラベルである。 (証明終わり)

例 7 L_2 の垂直要素 $\{100, 200, dash\}$ の代表ラベル集合は $\{dash\}$ であり、L_2 の代表ラベル集合 $\{dash, hurdles, high\}$ の部分集合である。

階層構造の性質より、1 つのラベルが複数の親を持つことはない。よって、2 つの垂直分割にまたがって同じラベルが属している場合、各垂直分割の根のラベルは上下関係を持っている。垂直分割は下位のラベルをすべて被覆しているため、一方の垂直分割はもう一方の垂直分割の要素となっている。

補題 3 ラベル集合 L の垂直要素 L_1, L_2 で $L_1 \cap L_2 \neq \emptyset$ ならば $L_1 \subseteq L_2$ または $L_2 \subseteq L_1$ である。
証明 $L_1 \cap L_2 \neq \emptyset$ であれば、$L_1 \cap L_2$ のラベル l が存在する。L_1, L_2 の最小被覆ラベルを l_{c1}, l_{c2} とすると、$l \prec l_{c1}, l \prec l_{c2}$ なので $l_{c1} \preceq l_{c2}$ または $l_{c2} \preceq l_{c1}$ であり、$L_1 \subseteq L_2$ または $L_2 \subseteq L_1$ である。 (証明終わり)

例 8 L_2 の垂直要素 $\{100, 200, dash\}$ と $\{100, 200, dash, hurdles\}$ において、$\{100, 200, dash\}$ が共通なラベルである。このとき、$\{100, 200, dash\}$ は $\{100, 200, dash, hurdles\}$ の部分集合である。

各垂直要素の根は元のラベル集合で下位のラベルであったラベルを被覆している。つまり分割を行っても、下位のラベルは各代表ラベルを根とするラベル集合に被覆されており、その代表ラベルよりも上位のラベルを根とした場合でも、その上位のラベルを根とするラベル集合に被覆されている。

補題 4 ラベル集合 L の垂直要素 L_1, L_2 で $L_2 \subsetneq L_1$ のとき、L_2 は L_1 の垂直要素である。

証明　L_2 は L の垂直要素なので、$L-L_2$ 中のラベルは L_2 中のラベルと上下関係がない。$L_1 \subsetneq L$ なので L_2 中のラベルは L_1-L_2 中のラベルと上下関係がない。$L_2 \subsetneq L_1$ なので、L_2 を要素とする L_1 の垂直分割が存在するため、L_2 は L_1 の垂直要素である。　(証明終わり)

例 9　L_2 の垂直要素 $\{100, 200, dash, hurdles\}, \{100, 200, dash\}$ において、$\{100, 200, dash\}$ は $\{100, 200, dash, hurdles\}$ の部分集合であり、垂直要素である。

5. 最小垂直分割

分割を行うことで、複数の流行を同一の流行ととらえずにすむ。しかし、過度に分割を繰り返した場合、分割後のラベル集合が小さくなりすぎてしまい、大きく流行をとらえることができない。本節では、垂直分割を行う際に、過度に分割を行わないよう検討する。

垂直分割を行うとき、分割を繰り返すと最終的にはラベル集合内の代表ラベルが 1 つになる。このように分割を過剰にしてしまうと、あるラベル 1 つのデータ数の変化となってしまい大きく流行をとらえたとはいえない。よって、分割の結果であるラベル集合はできる限り大きい集合でなければならない。できる限り大きい集合であると、ある 2 つの集約ラベルの和集合が ρ で制限されていない。

定義 6　L の垂直分割 $L_1, L_2, ..., L_m$ は、$L_i\ (1 \leq i \leq m)$ が ρ で制限されており、$L_i \cup L_j\ (i \neq j)$ の最小被覆ラベルに被覆される垂直要素の和集合は ρ で制限されていないとき、ρ による最小垂直分割であるという。

L の垂直要素は L の部分集合なので、垂直要素の最小被覆ラベルと垂直要素中のラベルの距離は、L の最小被覆ラベルと L 中のラベルの距離よりも短くなる。よって、垂直分割の垂直要素について、垂直要素の根であるラベルとその垂直要素の各ラベルとの距離も短くなる。

補題 5　ラベル集合 L が ρ で制限されているとき、L の垂直要素も ρ で制限されている。
証明　L が ρ で制限されているとき、L の最小被覆ラベルと L の代表ラベルとの距離は ρ 以下である。補題 2 より L の垂直要素 L' の代表ラベル l'_r は L の代表ラベルでもあり、L の最小被覆ラベルと l'_r の距離は ρ 以下である。L' は L の部分集合なので、L' の代表ラベル l'_r は L の最小被覆ラベルの下位である。よって L の最小被覆ラベルと l'_r の距離は ρ 以下であり、L' は ρ で制限されている。 (証明終わり)

例 10　閾値 ρ が 2 であるとすると、ラベル集合 $L_1 = \{100, 200, dash, hurdles\}$ の代表ラベルは $dash, hurdles$ である。最小被覆ラベルと代表ラベルの距離は 1 であり、ρ で制限されている。このとき、L_1 の垂直要素 $\{100, 200, dash\}$ も ρ で制限されている。

最小垂直分割で求まる垂直要素について、2 つの垂直要素の和集合の最小被覆ラベルは、2 つの垂直要素の共通の祖先である。このとき、和集合が ρ で制限されていれば、最小な分割ではないため、和集合の最小被覆ラベルによるラベル集合の集合は ρ で制限されない。また、求まった垂直要素を分割すると、ρ で制限されるが垂直要素が細かすぎてしまう。

定理 1　ラベル集合 L の ρ による最小垂直分割は唯一である
証明　ラベル集合 L の異なる最小垂直分割 $\boldsymbol{L_1}$ と $\boldsymbol{L_2}$ が存在すれば、$L_1 \cap L_2 \neq \emptyset, L_1 \neq L_2$ となる $\boldsymbol{L_1}, \boldsymbol{L_2}$ の要素であるラベル集合 L_1, L_2 が存在する。L_1, L_2 は L の垂直要素なので補題 3 より $L_1 \subset L_2$ または $L_2 \subset L_1$ である。$L_2 \subset L_1$ のとき、補題 4 より L_2 は L_1 の垂直要素である。L_2 を要素とする L_1 の垂直分割を $\{L_2, L_{11}, L_{12}, ..., L_{1m}\}$ とする。L_1 は ρ で制限されているので、補題 5 より L_{1i} も ρ で制限されている。$L_1 - L_2$ 中のラベル l' を含む $\boldsymbol{L_2}$ のラベル集合を L' とすると、$L_1 \cap L' \neq \emptyset$ なので補題 3 より $L_1 \subseteq L'$ または $L' \subseteq L_1$ であり、$L_2 \subseteq L_1 - L'$ なので $L' \subseteq L_1$ である。L_1, L_2, L' はラベル集合 L の垂直要素であり、$L_2 \subset L_1, L' \subset L_1$ なので、補題 4 より L_2, L' は L_1 の垂直要素である。補題 1 より $L_2 \cup L'$ の最小被覆ラベルに被覆されるすべてのラベル L の集合は L_1 の垂直要素であり、L_1 は ρ で制限されているので、補題 5 より $L_2 \cup L'$ の最小被覆ラベルに被覆されるすべてのラベル L も ρ で制限されている。$\boldsymbol{L_2}$ の垂直要素 L_2 と L' の和集合の最小被覆ラベルに被覆されるすべてのラベルが ρ で制限されているので、$\boldsymbol{L_2}$ が最小垂直分割であることに矛盾する。$L_1 \subset L_2$ の時も同様。よって異なる最小垂直分割は存在しない。すなはち最小垂直分割は唯一である。　(証明終わり)

最小垂直分割の結果が複数存在すれば、どのような流行が起きていることに対して複数の見方があることとなる。最小垂直分割の結果が唯一であることは、流行を特定できたということと同義であり、定理 1 より必ず特定できる。

最小垂直分割を求めるアルゴリズムは入力がある属性のラベル集合であり、出力は最小垂直分割された複数のラベル集合となる。出力される分割後の集約ラベルを $l^s_1, l^s_2, ..., l^s_m$ とすると、各ラベル集合は互いに上下関係を持たないため、$l^s_p \not\prec l^s_q\ (p \neq q)$ である。アルゴリズムにおいて、垂直分割は最小被覆ラベルが ρ で制限されていない場合、最小被覆ラベルの子孫側のエッジを繰り返し分割していき、分割後の各ラベル集合が ρ で制限されるところで止める。ある代表ラベル l_r の直属の子孫のラベルを $l'^s_k (k = 1, 2, ..., m)$ で表す。またこの各子孫のラベルを被覆するラベル集合を L'_k で表す。

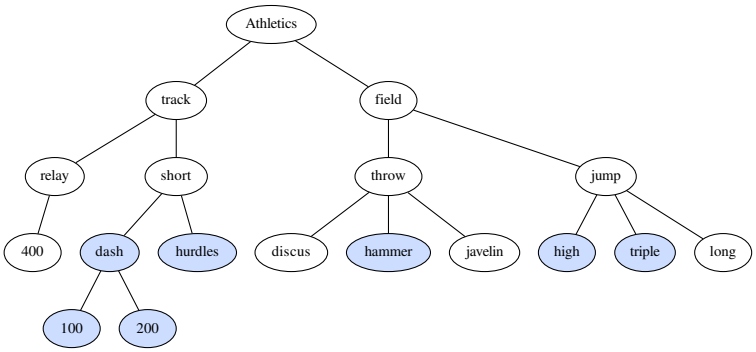

図 3 アルゴリズム

アルゴリズム 1 最小垂直分割

Input: $\boldsymbol{L} = \{L\}$
Output: 全てのラベル集合が ρ で制限されている L の垂直分割
while do $\boldsymbol{L}$ 中に ρ で制限されていないラベル集合 L' が存在
L' の最小被覆ラベルの子ラベル $l_1, l_2, ..., l_k$
$L_i = \{l \mid l \preceq l_i, l \in L'\}$
$\boldsymbol{L} = (\boldsymbol{L} - \{L'\}) \cup \{L_1, L_2, ..., L_k\}$
end while

アルゴリズム 1 は、ラベル集合 L の ρ による最小垂直分割を求める。L の垂直要素のうち、ρ で制限されていない垂直要素を分割し、新しい L の垂直分割とすることを繰り返す。垂直要素の分割は、その最小被覆ラベルの子ラベルを根とする部分木に含まれるラベル集合とする。

定理 2 アルゴリズム 1 で求まった分割は L の閾値 ρ による最小垂直分割である。
証明　各ステップのラベルの集合の集合を $\boldsymbol{L}_1, \boldsymbol{L}_2, ..., \boldsymbol{L}_n$ とする。$\boldsymbol{L}_1 = \{L\}$ であり、$\boldsymbol{L}_n$ はアルゴリズムの出力である。$\boldsymbol{L}_j$ が垂直分割であるとすると、$\boldsymbol{L}_j$ 中で選択された L' の分割 $L_1, L_2, ..., L_k$ は L' の垂直分割なので、$\boldsymbol{L}_{j+1}$ は L の垂直分割である。よってアルゴリズムで求まる分割は L の垂直分割である。

$\boldsymbol{L}_n$ 中の 2 つの垂直要素の和集合の最小被覆ラベルは、2 つの垂直要素の共通の祖先である。共通の祖先となっているラベルは分割する前の階層の根 l_0 と和集合の最小被覆ラベルの間に存在する。このラベルは分割の間に ρ で制限されないラベル集合の根となっているため、2 つの集約ラベルをまとめると ρ で制限されていない。よって 2 つの集約ラベル集合の和集合が ρ で制限されていないため、最小となる。　(証明終わり)

例 11 ラベル集合 L_3 を閾値 ρ を 1 として、アルゴリズムに基づいて分割を行う (図 3)。$L_3 = \{100, 200, dash, hurdles, hammer, high, triple\}$ であるとき最小被覆ラベルは $Athletics$ である。最小被覆ラベルと L_3 のラベル $\{dash, hurdles, hammer, high, triple\}$ との距離が 1 より大きく $Athletics$ で集約できない。

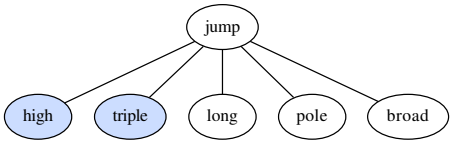

図 4 今後処理を考えるべきモデル

L_3 の最小被覆ラベル $Athletics$ の子ラベルは $track, field$ であり、分割を行うと、$track$ を根とする部分木に含まれるラベルは $L_4 = \{100, 200, dash, hurdles\}$ となり、$field$ を根とする部分木に含まれるラベルは $L_5 = \{hammer, high, triple\}$ となる。L_4 は最小被覆ラベルが $short$ であり、代表ラベルは $\{dash, hurdles\}$ である。代表ラベルと最小被覆ラベルの距離が 1 であり閾値以内であるため、L_4 は ρ で制限されており、$short$ は集約ラベルとなる。L_5 は最小被覆ラベルが $field$ であり、代表ラベルは $\{hammer, high, triple\}$ である。このとき、最小被覆ラベルと代表ラベルの距離が閾値 1 より大きいため分割を再び行う。
L_5 の最小被覆ラベルは $field$ であり、その子ラベルは $throw, jump$ である。$throw$ を根とする部分木に含まれるラベルは $L_6 = \{hammer\}$ となり、$jump$ を根とする部分木に含まれるラベルは $L_7 = \{high, triple\}$ となる。L_6 はラベル 1 つであるため $hammer$ が最小被覆ラベルかつ代表ラベルであり、$hammer$ が集約ラベルとなる。L_7 は最小被覆ラベルが $jump$ であり、代表ラベルが $\{high, triple\}$ である。L_7 の最小被覆ラベルと代表ラベルの距離は 1 であり、ρ で制限されているため $jump$ が集約ラベルとなる。よって、アルゴリズムによるラベル集合 L の最小垂直分割は $L_4 = \{100, 200, dash, hurdles\}, L_6 = \{hammer\}, L_7 = \{high, triple\}$ となる。

6. おわりに

本稿では、階層構造を用いたデータ数の変化について検討した。階層構造を用いることで、データ数の変化を個別に扱うのではなく、複数の変化をまとめることで、大きく流行をとらえることを可能とした。また、ラベルを集約する際にデータ数が増加しているラベルが存在する場合は分

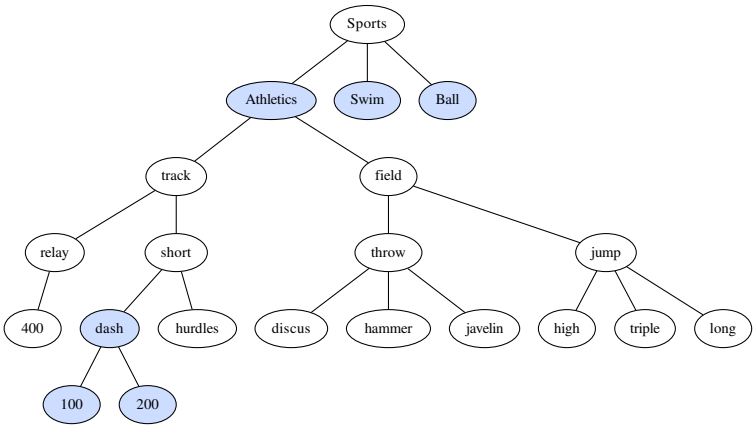

図 5 水平に流行をとらえるべき例

割を行うことで集約を可能とした。この分割の結果は唯一であり、流行を特定できることを示した。

今後はデータ数の変化とそれに応じた集約の手法を考える必要がある。本稿では、議論を簡単にするために共通な祖先が下位のラベルをすべて被覆することとした。しかし、あるラベルの子ラベルに極端にデータ数が多いラベルと、極端にデータ数が少ないラベルが共存し、それぞれ閾値を満たしていた場合に、それを共通の祖先でまとめることが正しいかどうかを検討する必要がある。

本稿に基づいて処理を行うと、閾値で制限されていると共通な祖先で集約するが、制限されていても集約すべきでない場合がある。たとえば、図 4 は $high, triple$ のデータ数が増加しており、最小被覆ラベルが $jump$ であるが、変化なしのラベルが $long, pole, vault$ と複数存在する。このように変化なしが混在してるケースにおいて、それらを集約させることが適当であるかを検討しなければならない。

また、データ数が増加したラベルと減少したラベルが混在した場合も存在する。たとえば、図 4 の $long, pole, vault$ のデータ数が減少していた場合、$high, triple$ を最小被覆ラベルで集約すると、データ数が減少しているものを含むため、集約すべきでない。このように、データ数が増加したラベルと減少したラベルが混在するケースは現実的に存在するため、検討しなければならない。

さらに、図 5 のような場合、$Sports$ が流行しており、それとは別で $dash$ が流行しているととらえるべきである。このように水平に流行をとらえるべきである場合の処理も考える必要がある。

また、本研究では 1 つの階層を想定したが、複数の階層の組み合わせについても検討する。複数の階層を組み合わせることで、属性をまたぎ流行をとらえることを可能とする。属性をまたいで流行をとらえるとき、属性同士の関連性などを考慮に入れることで精度の高い流行がとらえることができる。

参考文献

[1] Ema Kusen, and Mark Stermbeck: Politics, sentiments, and misinformation: An analysis of the Twitter discussion on the 2016 Austrian Presidential Elections, *Online Social Networks and Media*,Vol. 5, pp. 37-50, 2018

[2] Lisa Branz, and Patricia Brockmann: Sentiment Analysis of Twitter Data: Towards Filtering, Analyzing and Interpreting Social Network Data, *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems*, pp. 238-241, 2018

[3] Machiko Toyoda, Yasushi Sakurai, and Yoshiharu Ishikawa: Pattern discovery in data streams under the time warping distance, *The VLDB Journal June 2013*, Vol. 22, Issue 3, pp. 295–318, 2013

[4] Yizhou Sun , Kunqing Xie, Ning Liu, Shuicheng Yan, Benyu Zhang, and Zheng Chen: Causal relation of queries from temporal logs, *Proceedings of the 16th international conference on World Wide Web*, pp. 1141–1142, 2007

[5] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang: Quantifying Political Leaning from Tweets, Retweets, and Retweeters, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 8, 2016

[6] Hoonji Yang, Jiaofei Zhong, Dongsoo Ha, and Heekuck Oh: Rumor Propagation Detection System in Social Network Services, *International Conference on Computational Social Networks 2016 Computational Social Networks*, pp. 86-98, 2016

[7] Guannan Liu, Yanjie Fu, Tong Xu, Hui Xiong, and Guoqing Chen: Discovering Temporal Retweeting Patterns for Social Media Marketing Campaigns, *IEEE International Conference on Data Mining*, 2014

[8] John Scott: *Social Network Analysis*, 1992

[9] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong: Analysis of topological characteristics of huge online social networking services, *WWW '07 Proceedings of the 16th international conference on World Wide Web*, pp. 835-844, 2007

[10] Luiz Augusto, and Anmol Bhasin: Beyond friendship: the art, science and applications of recommending people to people in social networks, *ACM conference on Recommender systems*, pp. 495-496, 2013

[11] Luiz Pizzato, Tomasz Rej, Joshua Akehurst, Irena Koprinska, Kalina Yacef, and Judy Kay: Recommending people to people: the nature of reciprocal recommenders with a case study in online dating, *User Modeling and*

User-AdaptedInteraction, pp. 1–42, 2012

[12] Jeremy Ginsberg, Matthew Mohebbi, Rajan Patel, Lynnette Brammer, Mark Smolinski, and Larry Brilliant: Detecting in fluenza epidemics using search engine query data, *Nature*, Vol. 457, pp. 1012–1014, 2009.

[13] Sander Wozniak, Michael Rossberg, Guenter Schaefer: Towards trustworthy mobile social networking services for disaster response, *2013 IEEE International Conference on Pervasive Computing and Communications Workshops*, 2013