時空間情報に基づくイベント情報の集約システムの開発

廖　宸一[1]　梶　克彦[1]　廣井　慧[2]　河口　信夫[1,2]

概要：本研究は、時空間情報に基づくイベント情報の集約システム Event.Locky を開発する。Event.Locky はインターネット上で作成された大量のイベント情報を収集、集約、可視化するための Web サービスを提供するシステムである。本システムはさまざまな端末に対してデバイス環境適応型の Web フレームワークを採用し、いつでもどこでも使用可能な設計とする。イベント情報の集約に有益であるカテゴリ付けを活用し、機械学習の手法を用いて、カテゴリの付与されていないイベント情報のデータリソースに対し、カテゴリを推定するアルゴリズムを開発する。推定アルゴリズムを用いてイベント情報からカテゴリを推定したところ、アルゴリズムの標本カバレッジが 98.38%、適合率が 94.11%、再現率が 95.30%、F_1 値が 94.70%となり、高精度な推定が確認できた。また、実験からイベント情報の信頼性向上やカテゴリのノイズへの対処が必要であることがわかった。現在、本システムはインターネット上で公開しており、今後はユーザからのフィードバックを得てユーザビリティの向上を図る。

Development of Event Information Summarization System Based on Spatio-Temporal-Information

Chenyi Liao[1]　Katsuhiko Kaji[1]　Kei Hiroi[2]　Nobuo Kawaguchi[1,2]

1. はじめに

近年、インターネット上でのイベント情報が大量に作成されるようになった。伝統的な紙媒体の広告やポスターに比べると、情報量が大幅に増大するため、ユーザが大量のイベント情報から必要な知識を効率的に入手するのは困難である。そのため、それらのイベント情報の中からユーザに必要な知識や要約を取り出す知識処理技術 [1] が重要になりつつある。「イベント」は、「映画」や「交流会」のような催し物や行事を意味する [2]。

情報集約や大量のデータから知識を抽出するシステムは多く存在する。一円らの研究 [3] は複数のデータリソースから POI（Point of Interesting）情報を集約し、Linked Open Data を用いて集約されたデータを配布することを提唱した。Chen らの研究 [4] は自然言語処理の手法を用いてオンラインデータリソースから料理名の抽出や推薦を目指す。新聞記事の集約システム [5][6] も多く存在している。しかし、POI 情報や料理は時間によって変化しない情報が多い。新聞記事は時空間情報に依存するが、主に「過去」の出来事に注目する。災害情報の処理は時空間情報に依存する。発生した災害情報への処理によって意思決定を支援する研究 [7] があるが、新聞記事の集約システムのように、「過去」の出来事を扱う。時空間情報に依存し、将来の出来事（行う予定であるイベント）に注目するイベント情報の集約システムは少ない。そのため本研究は、時空間情報に基づくイベント情報を集約するシステムを開発する。イベントの行われる場所という空間情報および日時という時間情報はユーザに必要な知識の 1 つである。本研究で開発した Event.Locky[*1] はイベントサイトの提供した API を用い、大量のイベント情報を収集する。収集したイベント情報を空間情報と時間情報によって集約し、それらの情報を地図上に可視化するための Web サービスを提供するシステムである 。ユーザはイベント情報のリストを閲覧するだけでなく、地図上に表示される情報から、空間性が実感できる。そして、ユーザがイベントの期間を設定すると、設定された期間内のイベント情報が表示される。空間情報

[1] 名古屋大学工学研究科
Graduate School of Engineering, Nagoya University
[2] 名古屋大学 未来社会創造機構
Institute of Innovation for Future Society, Nagoya University

[*1] http://event.locky.jp/

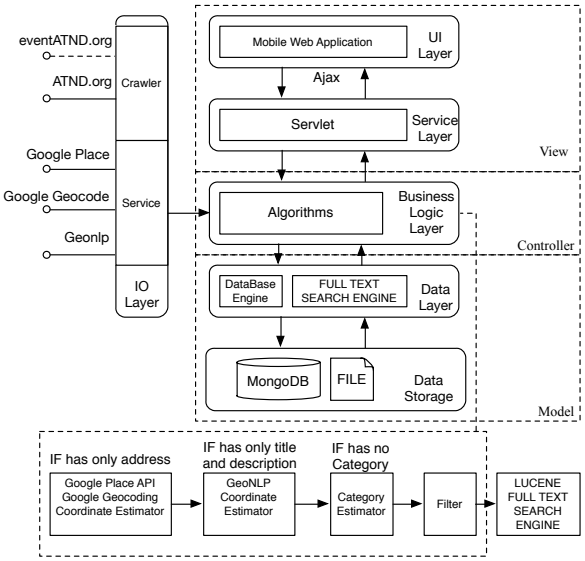

図 1　システムの構造

と時間情報との組み合わせの「時空間情報」も見ることができる。その時空間情報により、統計的な数値表現（地域別または時期別のイベント数など）ができる。

カテゴリ付けはイベント情報を集約する手法の 1 つであるが、イベント情報には、カテゴリの付与されていない Web サービスが多く存在する。そのため、カテゴリのないイベント情報に対してカテゴリ付けを行うためのカテゴリ推定アルゴリズムを提案する。

2. Event.Locky

2.1 Event.Locky の構造

本研究で開発する Event.Locky はインターネット上で作成された大量のイベント情報を収集、集約、可視化するための Web サービスを提供するシステムである。イベント情報はオンラインのイベント情報サービス API から収集され、データベースに保存される。いろいろな機能を実現するため、さまざまなモジュールを開発する。

各モジュールを協調的に行わせるため、MVC 構造 [8] を用いてシステムの構造を設計した。MVC は、アプリケーションソフトウェアの内部データをユーザが直接参照・編集する情報から分離するためのデザインパターンである。図 1 に示すように、MVC 構造から 5 つのレイヤを設計した。それぞれは、UI レイヤ、サービスレイヤ、ビジネスロジックレイヤ、データレイヤおよび IO レイヤである。データベースは NoSQL の MongoDB[11] を採用した。E-R モジュールを OPP モジュールに変換する必要はないため、従来のエンティティレイヤは採用していない。

まず、イベント情報収集のプロセスを説明する。イベント Web サービス API[12] から XML または JSON の形式で構造化データを配布している。IO レイヤのクローラは 4 時間おきに各イベント Web サービスの API からイベントデータの増分を取得する。RSS 標準 [13] を参考に、イベントのデータ構造を定義した。イベント情報はタイトル（title）、説明文（description）、開催時間（starttime）をイベント情報の基本的なデータ構造とする。その 3 つのデータフィールドを持つイベントデータが利用可能かを判別する。取得したイベントデータをビジネスロジックレイヤに入力してデータ構造を整理する。一般的に、イベント開催地の位置情報はイベント Web サービスの API から経緯度の形式で取られる。経緯度情報のないイベント情報は GeoNLP[14]、Google Place API[15] または Google Geocoding[16] を利用してイベント情報の住所から、経緯度情報を推定する。一方、本質的に位置情報を持たないイベント情報もある。例えば、オンラインイベントや場所未定のイベントなどは、「場所なしイベント」というリストで表示される。多くのイベント情報はカテゴリが付いているが、カテゴリの付かないイベント情報にとって、提案するカテゴリの推定アルゴリズムを用いてカテゴリを推定する。イベント情報を効率的に検索するために、Apache Lucene[17] 全文検索エンジンを採用した。整備されたイベ

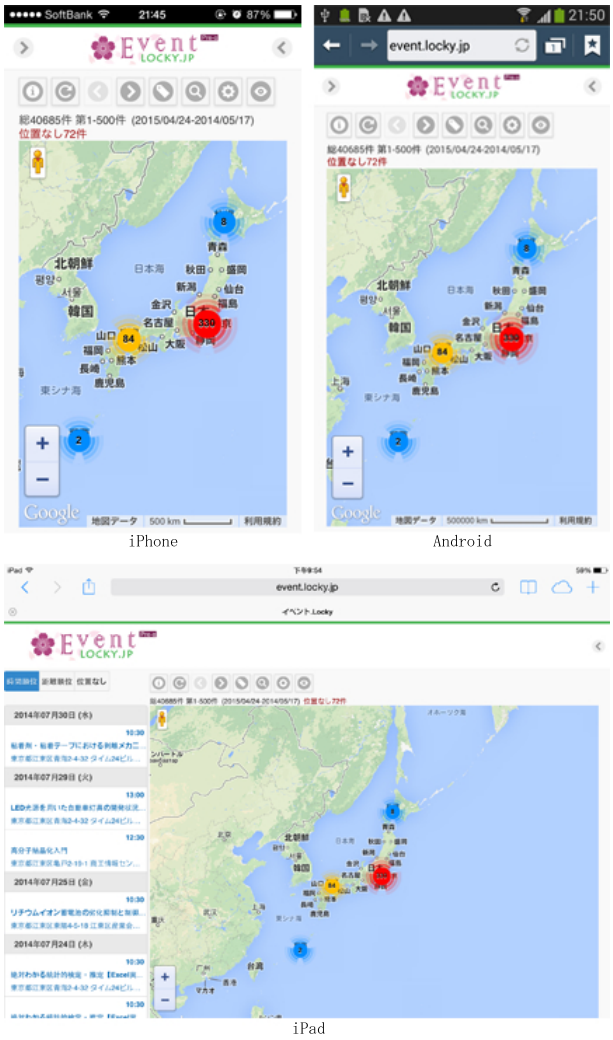

図 2　iPhone、Android、iPad でホームページの表示例

図 3　時間によってイベント情報を絞り込む例

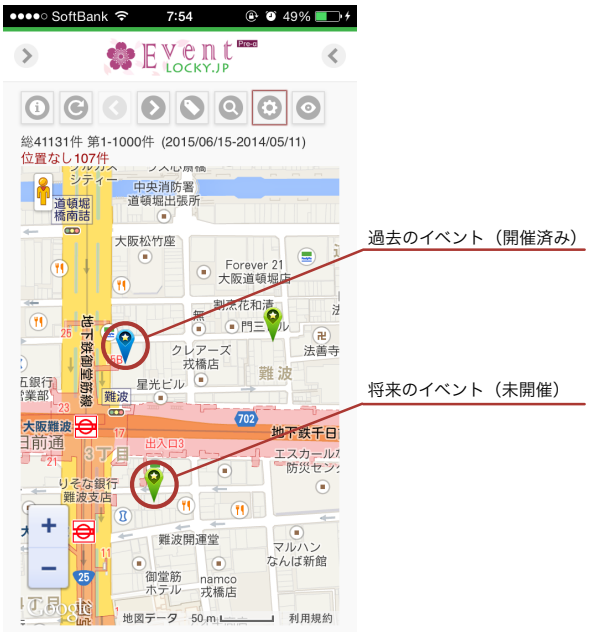

図 4　時間によってイベント情報を絞り込む例

ント情報で全文検索を用いてインデックスを作り、データベースに保存する。

データレイヤから、ビジネスロジックレイヤを経由し、サービスレイヤで HTML データに変換し、UI レイヤでイベント情報を可視化する。近年、スマートフォンを中心としたモバイルインターネットの発展とともに、ユビキタスコンピューティング [18] への利用が増えつつある。図 2 に示すように、デバイス環境適応型の Web フレームワークを採用して本システムの可視化インターフェースを開発した。HTML5 及び CSS3 の特性により、Web アプリケーションは端末によって表示レイアウトが自動的に変換できる。ユーザビリティを改善するため、クリエイト（ブラウザ）とサーバの間の通信は AJAX を採用した。将来的には、サービスレイヤを用い、集約されたイベントデータを API で配布する機能の開発を見込んでいる。

2.2　時間情報に基づくイベント情報の可視化

イベント情報は時間情報に緊密に依存しており、今現在の時点によって過去のイベントと将来のイベントに大きく分けられる。さらに、一般的なユーザ（データアナリストを除く）にとって、過去のイベントよりも将来のイベントの方が重要である。

そこで、時間によりイベント情報を絞り込む機能を開発する。図 3 に示すように、ユーザはこの画面で過去または将来のイベントが表示されるかどうかを設定する。また、地図上に過去または将来のイベント情報は別々のアイコンで表示される。図 4 に示すように、過去のイベントはブルーのアイコンで表示され、将来のイベントはグリーンのアイコンで表示される。図 5 に示すように、イベント情報は時間順でソートしたリストで表示する。位置情報は 2 段

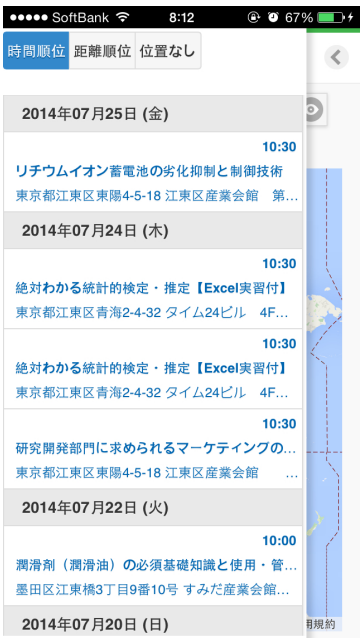

図 5　時間順位で表示されるイベント情報リスト

階で表示する。日ごとのサブリストのうちに、時間ごとの要素がある。

イベント情報の中には、図 6 に示すように、時間情報しか付かない（位置情報なし）イベント情報もある。それらのイベント情報は本質的に位置情報が付けられない場合が多い。例えば、オンラインイベントや場所未定のイベントなどである。さらに、信頼性の低いイベント情報も多く存在する。このようなイベント情報は「位置なし」というタブ内で、時間順位のリストを用いて表示する。

以上で述べた 4 つが本研究で用いた時間情報によるイベント情報の可視化手法である。もちろん、その他の可視化

図 6　位置情報なしのイベント

方法もある。例えば、現時点を中心値として過去または将来へリストが伸びていく表示方法もイベント情報の可視化に役に有効かもしれない。将来的には、より有益な可視化手法を検討する予定である。そして、イベント情報の信頼性の向上が今後の課題となる。

2.3　空間情報に基づくイベント情報の可視化

イベントはある特定の場所で開催されることが多いため、イベント情報は空間情報にも依存する。空間情報の可視化により、ユーザはイベントの開催地を効率的に見つけられるようになる。地図でイベントの開催地を表示するのは可視化手法の 1 つである。

より高い広域な視点から見ると、大量のイベントのアイコンが同一の地域に集中することが多く、ユーザビリティが下がる。そのため、本研究では同一の地域に集中したイベントのクラスタリングを提案する。図 7 に示すように、markerclustererplus.js[9] を用い、同一の地域に集中したイベントはその数をアイコンで表示する。そのアイコンをクリックすると、地図はそのアイコンを中心としてより詳細な地図を展開する。

また、同一の期間に、同一の開催地（会場、飲食店など）で複数のイベントが開催されることがある。この場合は、図 8 に示すように、OverlappingMarkerSpiderfier[10] を用いてそれらのイベントはレッドのアイコンで表示される。そのレッドのアイコンをクリックすると、開催地を中心としてイベントのアイコン（グリーンまたはブルー）が周囲に展開される。それらのアイコンは日時が遅いほど中心点から離れた位置へ表示される。

モバイル端末は位置情報が頻繁に変化するため、リアルタイムにモバイル端末とイベント開催地の間の距離をユー

図 7　同一地域のイベントのクラスター化

図 8　同一場所のイベントアイコンの展開

ザに提供するのは有効である。図 9 に示すように、システムは現在の位置からイベントの開催地への距離を計算し、距離順位リストでイベント情報を表示する。

ユーザがイベントを選択する時に、現在地からの距離といった位置に関連する情報は考える要素の 1 つである。そのため、位置情報を用いてユーザにイベントを推薦することができる。将来的には、位置情報を用いたイベント情報の推薦システムを検討する予定である。

3.　イベントのカテゴリ推定

3.1　カテゴリ推定アルゴリズムへの紹介

従来では、イベント情報にカテゴリの付いたイベント

図 9　距離順位で表示されるイベント情報リスト

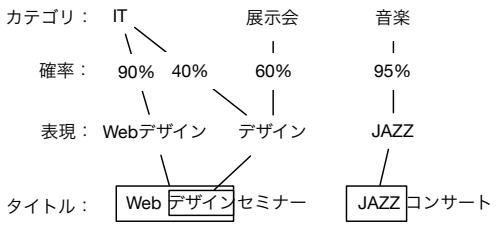

図 10　言語表現によってカテゴリを推定する例

Web サービスが多かった。eventAtnd は 18 種類のカテゴリを用いてイベントを区別しており、それぞれは、「グルメ」、「スクール」、「アート」、「お祭り」、「展示会」、「ショッピング」、「ファッション」、「エンタメ」、「映画」、「ボランティア」、「IT」、「ビジネス」、「交流会」、「スポーツ」、「ビューティ」、「くらし」、「演劇・演芸」、「音楽」であった。カテゴリはイベント情報の集約に非常に有益であり、ユーザがカテゴリを用いてイベントを効率的に検索することができる。しかし、2014 年 3 月 31 日に eventAtnd のサービスが終了してから、代わりの新しいサービス Atnd[12] はイベントにカテゴリを付与しない。そのため、カテゴリの付いていないイベント Web サービスにとって、そのイベント情報からカテゴリを推定し、推定されたカテゴリによって、イベント情報をグループ化する必要がある。

筆者らは機械学習の手法を用いてカテゴリの推定アルゴリズムを開発した。そのアルゴリズムは最近傍則 [19] に基づき、イベント情報（タイトルや説明文など）に含まれた特定の単語または句という言語表現とそれに該当するカテゴリとの相関関係によって、イベントに当たるカテゴリを判別するものである。

$$
\begin{aligned}
P(C|W) &= \frac{P(W|C)P(C)}{P(W)} \\
&= \frac{P(C,W)}{P(W)} \\
&= \frac{N(C \cap W)}{N(W)}
\end{aligned} \tag{1}
$$

図 10 に示すように、イベントのタイトルに含まれた「Web デザイン」や「デザイン」などの言語表現で、その言語表現に最近傍（確率の最大値）のカテゴリを選定する手法を提案する。こちらの確率は特定な言語表現が起きる場合、カテゴリが起きる事後確率を求めるために、ベイズ推定 [20] の手法を用い、言語表現に当たるカテゴリの起きる回数とその言語表現の起きる総回数の割合を求めた。式 1 に示すように、C がカテゴリ、W が言語表現、W が言語表現、$P(W)$ が言語表現の起きる確率、$P(C)$ がカテゴリの起きる確率、$P(W|C)$ がカテゴリの起きる場合、言語表現の起きる確率、$P(C|W)$ が言語表現の起きる場合、カテゴリの起きる確率、N が起きる回数を指す。例えば、タイトルに「デザイン」があるイベントは総 10 件ある（$N(\text{デザイン}) = 10$）。そのうち、4 件はカテゴリ「IT」が付いている（$N(\text{IT} \cap \text{デザイン}) = 10$）。そのため、タイトルの中に「Web」がある場合、イベントはカテゴリ「IT」になる確率は

$$
P(\text{IT}|\text{Web}) = \frac{N(\text{IT} \cap \text{デザイン}}{N(\text{デザイン})}) = \frac{4}{10} = 40\%
$$

になる。

$$
\begin{aligned}
P(C|W) &= \frac{P(W_1 \cdots W_i \cdots W_n|C)P(C)}{P(W_1 \cdots W_i \cdots W_n)} \\
&= \frac{P(C, W_1 \cdots W_i \cdots W_n)}{P(W_1 \cdots W_i \cdots W_n)} \\
&= \frac{N(C \cap W_1 \cdots \cap W_i \cdots \cap W_n)}{N(W_1 \cdots \cap W_i \cdots \cap W_n)}
\end{aligned} \tag{2}
$$

複合句や共起語を扱う言語表現がその確率に大きな影響を与える。図 10 に示したように、単に「デザイン」でカテゴリを判断する場合、カテゴリ「IT」に当たる確率は 40% になるが、「Web」と「デザイン」を複合句として共起する場合、カテゴリ「IT」の確率は 90% にあがり、判別境界が更に明確になるため、式 3 を利用する。文の中に複数個の言語表現が共起した場合、カテゴリの起きる確率を求める。その機能を実現するため、CYK アルゴリズム [21] を採用した。CYK（Cocke-Younger-Kasami）法は文脈自由文法で文字列を生成する方法の 1 つであり、構文木による Bottom-Up の構文方法である。

CYK 法で生成された CYK テーブルを図 11 に示す。元の CYK テーブルが図 11 の左側に示されたように、各形態素（単語）の組み合わせで、文の構成を分析するものである。本研究では、CYK 法を用いて形態素の各組み合わせを調査し、その形態素の組み合わせとカテゴリの相関を統

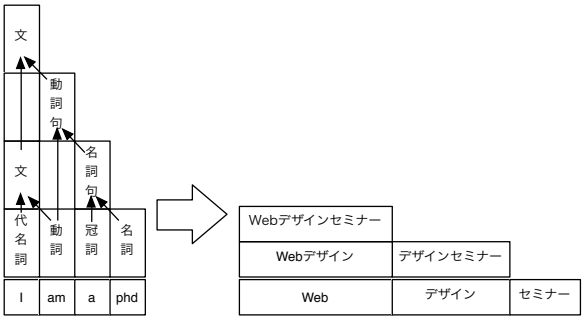

図 11 CYK テーブルの例

Webデザインセミナー
形態素解析

Web	名詞,固有名詞,組織,*,*,*
デザイン	名詞,サ変接続,*,*,*,*,デザイン,デザイン,デザイン
セミナー	名詞,一般,*,*,*,*,セミナー,セミナー,セミナー

図 12 MeCab を用いて形態素解析を行う例

計的に求める。

単語の場面から見ると、英語と違い、日本語文には単語ごとの分割がない。そのため、日本語文を処理する場合、形態素解析 [22] が必要である。形態素解析は文を意味を持つ最少単位の列に分割し、それぞれの品詞を判別する作業を指す。日本語の形態素解析エンジンにはいくつかの既存研究がある。本研究では、京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所の共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジン MeCab（和布蕪）[23] を採用している。MeCab を用いた形態素解析の実行例は図 12 に示されるように、各形態素の表層格、品詞、平仮名・片仮名・漢字変換などの情報が得られる。

3.2 実験で用いたイベント情報

eventAtnd がサービス終了する前、そのサービスから 16,063 件のイベント情報を収集した。収集された 16,063 件のカテゴリ付けのイベント情報から、利用不可のイベント情報（カテゴリが空きまたは「その他」）を除き、利用可能なイベント情報 12,988 件を実験標本とした。そのうち、90%の実験標本を訓練標本（11,689 件）として、10%をテスト標本（1,299 件）として実験を行った。

$$C = \frac{N(POS)}{N(S)}$$
$$P = \frac{N(E \cap I)}{N(E)}$$
$$R = \frac{N(E \cap I)}{N(I)}$$

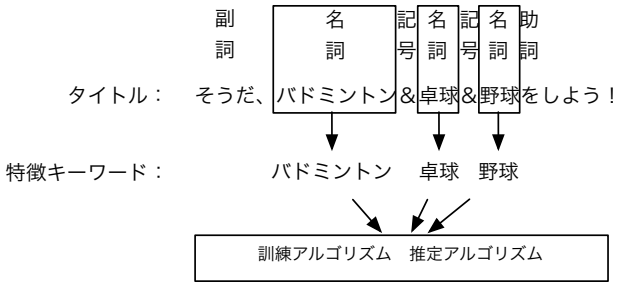

図 13 特徴キーワードの選択例

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} = \frac{2PR}{P+R} \qquad (\beta = 1) \qquad (3)$$

式 3 に示すように、カテゴリ推定アルゴリズムの評価は「標本カバレッジ（C)」、「適合率（P)」、「再現率（R)」と「F_1 値（F_1)」の 4 つの指標 [24] で行う。「標本カバレッジ（C)」は標本に選定された品詞に含まれた標本（C）の割合を指す。「適合率（P)」は推定された結果（$N(E)$）のうち正しく推定された結果（$N(E \cap I)$）の割合を指す。「再現率（R)」は元のカテゴリ（$N(I)$）のうち正しく推定された結果（$N(E \cap I)$）の割合を指す。「F 値（F_β)」は適合率と再現率の調和平均であり、高ければ、性能が良いことを意味する。一般的に β は 1 に設定され、F_1 値になる。

3.3 特徴キーワードの選択

訓練アルゴリズムまたは推定アルゴリズムに特徴量を入力する前、適当な形態素を選択する必要がある。例えば、イベントのタイトル「そうだ、バドミントン&卓球をしよう！」にとって、「バドミントン」・「卓球」などの名詞はカテゴリ「スポーツ」に相関が強いが、「そう」・「&」・「を」・「しよう」などの副詞や助動詞、記号、助詞はカテゴリに相関が弱い。本文では、それらのカテゴリに相関の強い形態素を特徴キーワードと呼ぶ。特徴キーワードの選択はシステムの効率と予測の精度に大きな影響を与える。特徴キーワードを少なく選択した場合、標本カバレッジが不足し、精度が低くなる。特徴キーワードを多く選択した場合、計算量と負荷が増え、システム効率が悪くなり、ノイズが大きくなってしまう。図 13 に示すように、日本語文法の視点から見ると、形態素の品詞を用いて特徴キーワードを選択することが可能である。特徴キーワードを選択する場合、一般的にはすべての品詞の組み合わせを試み、一番 F_1 値の高い品詞を選択する手法がある。本研究は異なる判別境界（0.00-0.99）で各組み合わせの F_1 値がピークになった時、標本カバレッジ・判別境界・適合率・再現率と F_1 値を計算した。表 1 は F_1 値が 0.24 より高い品詞とその組み合わせを示した。表 1 の上部は単一の品詞を選択した結果であり、下部は複数の品詞の組み合わせを選択した結果である。

その結果、「名詞一般」・「固有名詞組織」・「名詞サ変接続」での品詞の組み合わせを特徴キーワードとした場合、F_1 値が一番高いという結果が得られた。98.28%のカバレッジは

表 1　品詞別の実験結果

品詞	総標本カバレッジ	判別境界	適合率	再現率	F_1 値
NG（名詞一般）	91.16%	0.72	56.55%	74.52%	61.71%
SS（記号空白）	28.00%	0.30	46.34%	46.19%	46.24%
NSN（名詞接尾助数詞）	42.40%	0.63	35.98%	35.76%	35.81%
NN（名詞数）	62.02%	0.42	26.54%	49.69%	33.36%
NPO（名詞固有名詞組織）	35.12%	0.18	31.65%	33.45%	32.08%
NPPG（名詞固有名詞地域）	35.13%	0.12	21.47%	48.85%	29.08%
NS（名詞サ変接続）	72.75%	0.08	15.38%	68.48%	24.62%
NG ∪ SS	96.69%	0.80	64.48%	84.22%	69.04%
NG ∪ SS ∪ NSN	97.14%	0.80	64.63%	84.33%	69.16%
NG ∪ SS ∪ NSN ∪ NN	97.44%	0.80	55.56%	85.07%	62.55%
NG ∪ SS ∪ NSN ∪ NPO	98.68%	0.80	63.28%	84.99%	68.48%
NG ∪ SS ∪ NPO	98.50%	0.80	63.28%	85.06%	68.51%
NG ∪ NPO	98.31%	0.72	6690%	84.95%	72.09%
NG ∪ NPO ∪ NS	98.38%	0.72	68.07%	86.91%	73.37%

表 2　訓練アルゴリズム

Algorithm 1 Training Algorithm

1 begin initialize $V_{[len][len]}$ as a string array
2 i ← 0
3 do $i \leftarrow i+1$
4 $V_{i1} \leftarrow \{A|A \leftarrow x_i\}$
5 until $i = n$
6 $i \leftarrow 0$
7 $j \leftarrow 0$
8 $k \leftarrow 0$
9 $V_{ji} \leftarrow V_{jk}$ connects with $V_{[k+j+1][i-k-1]}$
10 until $k<i$
11 Store(C, V_{ji}, categories++)
12 until $j<len-i$
13 until $i<len$

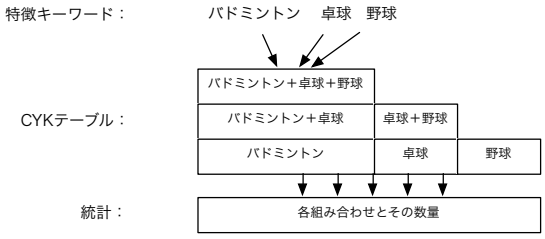

図 14　訓練アルゴリズムの例

システムの許容範囲内であると考える。そのため、本システムでは「名詞一般」・「固有名詞組織」・「名詞サ変接続」の組み合わせを用いてカテゴリ推定アルゴリズムを設計した。

3.4　訓練アルゴリズムの設計

訓練アルゴリズムは CYK 法を用い、入力された特徴キーワードの各組み合わせとそれらに当たるカテゴリを統計的に記録するアルゴリズムである（図 14）。表 2 に示すように、9 行目の書き換え規則で複合句や共起語を連結し、11 行目で図 15 に示すように、各特徴キーワードとそのカテゴリの数量をデータベースに記録する。

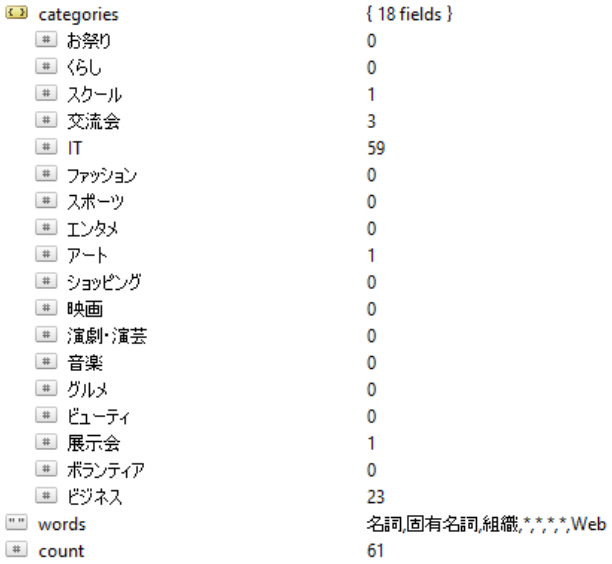

図 15　統計された特徴キーワードとそのカテゴリのデータ

3.5　各カテゴリの判別境界

4.2 節に述べた品詞の組み合わせの実験ですべてのカテゴリにとって統一の判別境界を設定してその F_1 値を求めたが、実際に各カテゴリの判別はそれぞれの 2 値分類問題であるから、カテゴリによって別々の判別境界を与える必要がある。そのため、各カテゴリの判別境界を求める必要がある。表 3 示したように、F_1 値がピークになった場合、各カテゴリの F_1 値を求めた。

3.6　推定アルゴリズムの設計

推定アルゴリズムは訓練アルゴリズムの逆過程とは言える。まず、18 種類のカテゴリの確率を 0.0 に初期化する。次に、CKY 法を用いて特徴キーワードの各組み合わせをデータベースに記録された訓練データと比較し、各カテゴリの確率を計算する。例えば、図 15 に示したように、タイトルに特徴キーワード「Web」のある場合、カテゴリ「IT」

表 3　カテゴリ別の判別境界実験

カテゴリ	標本数	判別境界	適合率	再現率	F_1
スクール	1,218	0.94-0.95	79.47%	92.54%	85.51%
ファッション	131	0.84-0.99	77.86%	93.58%	85.00%
エンタメ	3,732	0.98	94.76%	98.31%	96.50%
音楽	1,044	0.98-0.99	88.86%	95.50%	92.06%
スポーツ	750	0.92-0.96	79.31%	95.07%	86.48%
アート	533	0.99	89.96%	94.18%	92.03%
ボランティア	254	0.97-0.99	83.83%	87.80%	85.77%
グルメ	78	0.93	73.27%	94.87%	82.68%
交流会	4,361	0.99	87.91%	93.19%	90.47%
ビジネス	2,361	0.98-0.99	69.47%	93.56%	79.73%
お祭り	301	0.86-0.99	76.81%	84.72%	80.57%
IT	1,727	0.98-0.99	62.02%	92.47%	74.24%
くらし	414	0.83-0.85	71.93%	89.13%	79.61%
ショッピング	88	0.89-0.99	90.41%	75.00%	81.99%
ビューティ	227	0.91	81.15%	92.95%	86.65%
展示会	105	0.77	88.00%	83.81%	85.85%
映画	78	0.92	71.84%	94.87%	81.77%
演劇・演芸	72	0.89-0.99	82.76%	100.00%	90.57%

図 16　各カテゴリの確率の例

の確率は

$$P(\mathrm{IT}|\mathrm{Web}) = \frac{N(\mathrm{IT} \cap \mathrm{Web})}{N(\mathrm{Web})} = \frac{59}{61} = 96.72\%$$

になる。次に、最近傍則により、計算されたカテゴリの確率は従来のそのカテゴリの確率（初期は 0.0）より大きければ、確率値をより大きい方に設定する。その結果、図 16 に示すように、各カテゴリの確率が得られた。最後、表 3 に示す各カテゴリの判別境界により、判別境界より確率の高いカテゴリをイベントのカテゴリに判別する。

3.7　実験結果と実装

実験の結果、カテゴリ推定アルゴリズムの適合率が 94.11%に、再現率が 95.30%に、F_1 値が 94.70%となった。現在、アルゴリズムをシステムで実装実験を行っている。

なお、2014 年 4 月 21 日にインタネット上でシステムの pre-alpha 版を公開した。同年 5 月 3 日までイベント情報を 40,634 件収集した。

表 4　推定アルゴリズム

Algorithm 1 Training Algorithm

```
1   begin initialize V_[len][len] as a string array
2    i ← 0
3    do i ← i + 1
4     V_i1 ← {A|A ← x_i}
5    until i = n
6    i ← 0
7     j ← 0
8      k ← 0
9       V_ji ← V_jk connects with V_[k+j+1][i−k−1]
10     until k<i
10     if P(V)<Probability(C, V_ji, categories)
11     then P(V) ← Probability(C, V_ji, categories)
12    until j<len − i
13   until i<len
```

4.　おわりに

本研究は、イベントに関するカテゴリ推定アルゴリズムを提案し、時空間情報に基づくイベント情報の集約システム Event.Locky を開発した。Event.Locky はインターネット上で作成された大量のイベント情報を収集、集約、可視化するための Web サービスを提供するシステムである。本システムはさまざまな端末に対してデバイス環境適応型の Web フレームワークを採用し、いつでもどこでも使えるようにした。イベント情報の集約に有益であるカテゴリ付けに着目し、カテゴリが付かないイベント情報のデータリソースに対し、機械学習の手法を用いてカテゴリを推定する。開発したシステムを用いてカテゴリ推定の評価実験を行ったところ、アルゴリズムのカバレッジが 98.38%、適合率が 94.11%、再現率が 95.30%、F_1 値が 94.70%に達成した。そして、イベント情報の収集について、以下の 2 つの課題が得られた。

課題の 1 つ目にイベント情報の信頼性がある。収集されたイベント情報の考察を通じ、一部のイベント情報の信頼性が低いことがわかった。あるデータの登録者が勝手に登録したイベント情報には、「テストイベント」のような偽のイベント情報がある。それらの信頼性の低いイベント情報を取り除く必要がある。今後は、時空間情報に基づいた異常検知の手法を用いて信頼性の低いイベント情報を検知する方法を検討する。

2 つ目の課題としてカテゴリ付けがあげられる。本研究で用いた 18 種類のカテゴリは eventAtnd の開発者が事前に設定したものであるため、あるカテゴリにとってのイベ

ント情報が非常に疎になる傾向がある。表 3 に示したように、標本数が最大のカテゴリは、最少のカテゴリの 60 倍以上になってしまう。そして、標本量が少ない場合、判別境界が不明確になる。標本量の少ないカテゴリにとって判別境界区間が大きすぎると、判別境界の閾値を設定しにくくなるという問題があり、推定精度に悪い影響を与える可能がある。疎性問題に対応するため、大量の訓練標本を使わなければならない。

一方、それらのカテゴリを設定する方法がユーザビリティに良いかどうかは分からない。あるカテゴリ（交流会）は他のカテゴリ（IT）を含むことが多く、明確に分類できないことがある。そのため実際にシステムを使う場合、既定されたカテゴリの可用性が問題となる。

イベント情報へのカテゴリ登録は、登録者によって行われるため、統一の基準でカテゴリ付けが行われない場合、そのイベント情報はノイズとなる。本稿の評価実験で、そのノイズが推定アルゴリズムの性能に影響を与えることがわかった（カテゴリ「IT」）。そのため、集約システムや推薦システムを開発すれば、ノイズの蓄積でシステム性能が低下する可能がある。さらに、カテゴリはあらかじめ規定されているため、扱えないイベント情報が存在する。本研究で使われた標本の 16,063 件の中、3,075 件（19.14%）が「その他」または「カテゴリなし」が付けられている。そのため、クラスタリングアルゴリズムを利用して自律的なイベントの分類を行うことで、既定カテゴリ付けのノイズに対応する。今後はシステムの改良とユーザビリティの向上を目指すすとともに、将来的には複数のイベントデータリソースからイベント情報の収集システムを改良する予定である。

謝辞

本研究の一部は、Microsoft Research Asia (MSRA) と，総務省戦略的情報通信研究開発推進事業（SCOPE）132306007 の助成をうけて実施された。

参考文献

[1] Ahamed, Syed Vickar, and Victor Bernard Lawrence. "Knowledge processing system employing confidence levels." U.S. Patent No. 5,809,493. 15 Sep. 1998.

[2] 新明解国語辞典 第七版

[3] Shinji Ichien, Katsuhiko Kaji, Nobuo Kawaguchi: Proposal of a Platform Integrating POI Information, ICMU, pp124-128, 2014.

[4] Weichang Chen, Katsuhiko Kaji, Nobuo Kawaguchi: Non-Local Dictionary Based Japanese Dish Names Recognition Using Multi-Feature CRF frame Online Reviews, WIMS, 2014.

[5] 地図新聞: 2014.05.03, http://www.mapnews.jp/.

[6] ニュースマップ世界: 2014.05.03, http://news.goo.ne.jp/world/map/.

[7] 村崎大輔, 藁科光徳, 小池英之, 荒川淳平, 上田真史, 竹内郁雄. (2009). 災害情報可視化システムの開発. 日本地震工学会論文集, 9(2), 2_88-2_101.

[8] Glenn E. Krasner,Stephen T. Pope: A cookbook for using the model-view controller user interface paradigm in Smalltalk-80. Journal of Object-Oriented Programming. pp26-49. Volume 1 Issue 3, Aug./Sept. 1988

[9] markerclustererplus: 2014.05.16, https://github.com/mahnunchik/markerclustererplus

[10] OverlappingMarkerSpiderfier: 2014.05.16, https://github.com/jawj/OverlappingMarkerSpiderfier

[11] MongoDB: 2014.05.07, http://www.mongodb.org/

[12] ATND API: 2014.05.07, http://api.atnd.org/

[13] The application / rss+xml Media Type. Notwork Working Group. May 22th 2006

[14] GeoNLP: 2014.05.07, https://geonlp.ex.nii.ac.jp/

[15] Google Places API: 2014.05.07, https://developers.google.com/places

[16] Google Geocoding API: 2014.05.07, https://developers.google.com/maps/documentation/geocoding

[17] Apache Lucene: 2014.05.07, http://lucene.apache.org/

[18] Pendyala, V.S., Shim, S.S.Y.: The Web as the Ubiquitous Computer. Computer 42(9) (2009).

[19] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." Information Theory, IEEE Transactions on 13.1 (1967): 21-27.

[20] Cheeseman, Peter. "A method of computing generalized Bayesian probability values for expert systems." Proceedings of the Eighth international joint conference on Artificial intelligence-Volume 1. Morgan Kaufmann Publishers Inc., 1983.

[21] Younger, Daniel H. "Recognition and parsing of context-free languages in time n^3." Information and control 10.2 (1967): 189-208.

[22] 工藤拓, 山本薫, 松本裕治. "Conditional Random Fields を用いた日本語形態素解析." 情処学 NL 研報 (2004): 161-13.

[23] Kudo, Taku. "MeCab: Yet another part-of-speech and morphological analyzer." http://mecab. sourceforge. net/ (2005).

[24] Buckland, Michael K., and Fredric C. Gey. "The relationship between recall and precision." JASIS 45.1 (1994): 12-19.