Vol.2009-DBS-149 No.22
2009/11/21

時空間情報を伴うテキストデータの要約システムの設計

山中　努[†1]　土方嘉徳[†1]　西田正吾[†1]

近年 Web 上で使用できる地図アプリケーションや GPS 機能が付いた携帯電話が普及しつつある．また twitter や場 log のように GPS 情報を付加して周囲の状況をテキストで送ることができるサービスが登場しつつある．これらにより時空間情報を伴うテキストデータが増加しつつある．この大量の情報をうまく活用できればイベント会場の管理者や災害時における自治体のオペレータのように，ある特定の地域の状況を把握する事が必要となる業務において，より迅速で正確な状況把握を実現する事ができると思われる．そこで本研究では時空間情報を伴う大量のテキストデータを業務上必要な観点から要約し，その情報を地図上で可視化するシステムを設計する．

Architectonics of a text-summarization system with temporal-spacial data

Tsutomu Yamanaka,[†1] Yoshinori Hijikata[†1]
and Shogo Nishida [†1]

Map applications on the web or mobile phones with GPS capabilities have become widely used. Furthermore, users can send surrounding circumstances via email with GPS information by using twitter. So text-data with temporal-spacial infomation is increasing. In business that they have to understand local conditions like venue administrator or operators of local governments in disaster, they can use this information for understanding circumstances faster and more accurately. In this study, we extract important information for this business from large amounts of text data with temporal-spatial information and design a system that visualizes the information on the map.

†1 大阪大学大学院 基礎工学研究科
Graduate School of Engineering Science, Osaka University

1. はじめに

近年 Web 上で使用できる地図アプリケーションや GPS 機能が付いた携帯電話が普及し，twitter[1] や場 log[2] のように電子メールに GPS 情報を付加して周囲の状況をテキストで送ることができるサービスが登場した．これに伴い時空間情報を伴うテキストデータが増加しつつある．今後大量に時空間情報を伴うテキストデータを取得できるようになれば，これをビジネスでも有効活用したいというニーズが発生すると思われる．例えばイベント会場の管理者が会場内の問題や混雑状況などを調べる場合や地震や津波などの災害時に自治体のオペレータが被災状況を調べる場合などにおいてである．そこで本研究では時空間情報を伴うテキストデータを利用して，業務向けの状況把握システムを開発する．

システムを設計する上で重要な事は，如何にイベント会場の管理者や自治体のオペレータなどの意思決定者が短時間で周辺の状況を把握できるかである．しかし時間，空間の情報をそのままテキスト形式で表示するだけでは情報を直感的に把握することはできない．そこで情報をユーザにとって分かりやすい形に要約し，可視化してやる必要がある．特に業務を対象とした場合，注目する情報の観点はある程度決まっており，さらに迅速に状況の変化に対処しなければならない．したがってこれらを考慮した上で情報を要約する必要がある．

経度・緯度の空間情報を属性として持つ情報を可視化表示する上で一般的な方法は，１つの情報につき１つのアイコンを対応させ，経度・緯度に基づいて地図上にアイコンをマッピングする事である．このような可視化表示は，Google マップ[3] や YAHOO!地図[4] などの Web 上の地図アプリケーションにも見られる．また自治体の災害対策本部にも該当区域の地図を張った机やそれを表示した大型スクリーンなどが設置されていることが多く，そこにオペレータが各地で起こった情報を書き込み状況把握を行っている．

よって本研究でもアイコンを用い地図上に一般市民から送られる情報を重畳表示させることを考える．しかし単純に１つの情報につき１つのアイコンを対応させるのでは問題が生じる．送られてくる情報が少ない場合や，ホテルやレストランの地図上での検索のように情報の重複が無いような場合にはこの手法で十分である．しかし類似した内容を多く含む情報が大量に投稿されるような場合に１つの情報につき１つのアイコンを対応させる手法を取ると重畳表示される情報が非常に多くなってしまう．またユーザは異なる情報を見たいと思い複数のアイコンをクリックしても同じ内容の情報を閲覧する事が多くなってしまう．この問題に対して解決策を示す必要がある．

以上により本研究では，業務上必要な観点から情報を要約し，情報数やシステム利用者が

調べたい内容に応じて地図上でインタラクティブに情報を表示できるようなシステムを設計する．

以下 2 章では関連研究を，3 章ではシステムの概要を説明する．4 章では具体的な実装方法を示す．5 章では我々の提案するシステムが十分に機能するかを実験により確かめ，6 章でまとめと今後の課題を述べる．

2. 関 連 研 究

時空間情報を伴うテキストデータを要約し，可視化する研究は本研究以外においてもいくつか行われている．倉島ら[5] はブログを対象とし，ブログの投稿時間から時間情報を，ブログのテキストデータから地名表現を抜き取ることにより空間情報を取得し，時空間情報を伴うテキストデータを取得している．そして時間情報とテキスト情報，空間情報の共起関係を相関ルール分析により発見し，ある地域での人々の行動を把握する事を可能にしている．可視化は地図上で行っており，地名を中心にして半透明な円形領域が広がり，その円形領域内に地名に関する数個の話題語が表示される．

また宮森ら[6] は携帯端末等により街の様々な場所で人々が感じた事をつぶやきのような短い発話として時空間情報と共に発信される状況を想定し，街の任意の時空間的な範囲における特徴的なイベントや状況を集約して・提示している．集約，提示方法としては以下の通りである．まず時間を週，または月単位で区切り，その時間範囲で区切られるデータを取得する．次に取得したデータを空間情報に基づいてクラスタリングする．この時各クラスタの中心位置 c と距離 r を得る．次にクラスタに含まれる発話テキストを形態素解析し，単語の出現頻度をもとにクラスタを一言で表すラベルを作成する．そして地図上に中心 c，半径 r の半透明な円を描き，円の中心にラベルを提示する．

これら２つの研究と我々の行う研究の相違点は，これら２つの研究が娯楽や旅先を探すといった日常の目的で使われる事を想定しているのに対し，我々の研究では業務目的で使われる事を想定している点である．3 章で詳しく説明するが，我々のシステムでは業務を決定する事によりより目的志向の要約・可視化手法を可能にしている．また前者の研究は時空間情報を伴うテキストデータが得られるもののブログを投稿する時間と実際にブログで書く内容を体験した時間とにずれがあり，また空間情報においても正確にテキストから割り出す事は困難である．すなわち情報発信時における詳細な空間情報と時間情報を必ず持っているか否かという点で我々の対象とするデータとは異なる．このため前者の研究では比較的広い地理的範囲での情報の要約が主となる．

3. システム構成

3.1 システムの特徴

本システムでは 1 章で述べたような業務上の性質と，大量の類似した情報を可視化する場合の問題点を考慮して以下の３つを特徴的な機能として考えた．

(1) 電子メールの自動分類
(2) 情報の統計的要約
(3) 情報可視化

まずこの３つの機能について，その機能の説明と，その機能が必要となる背景を述べる．

電子メールの自動分類

業務において注目する内容の観点はある程度決まっている．例えばイベント会場の管理者であるならば施設や人々に関する情報などを，災害時であるならば水害，火災，避難所，土砂崩れなどに関する情報などが注目する内容として挙げられる．この場合，注目する内容の観点から情報を検索できると目的の情報が探しやすくなると思われる．そこであらかじめカテゴリを注目する内容の観点から設定しておき，新たに入ってきた情報がカテゴリに属するか否かを分類する自動分類 (classification) を行う．自動分類の手法としては，１つのメールが必ず設定したカテゴリのうちの１つに属するようにする手法と，１つのメールが０以上の複数のカテゴリに属するようにする手法が考えられる．周辺状況をメールで送るという場合，１つのメールが複数のカテゴリに言及しているケース（例えば「風雨がひどく土砂崩れが起き，水害が発生しています.」など）あるいはどのカテゴリにも属さないようなケースも少なからずあると考えられるためここでは後者の手法を採用する．

情報の統計的要約

オペレータ（以降システム利用者をオペレータと呼ぶ）は期間・場所・カテゴリを選択して取り出したい情報の絞込みができるようにする．しかし絞り込み条件によっては類似した内容を多く含む情報が大量に出力される場合がある．そこで状況を把握を容易にするため，クラスタリングを行うことにより良く似た情報を１つにまとめて表示させる．

また現在起こっている問題を知りたいという要求がオペレータには存在する．如何に現在起きている問題を素早く察知し，対処できるかが損失を最小限に抑えるポイントになるからである．具体的には，上記のクラスタリングにより，似た情報の集合が特定されているので，その集合内における各情報の時間的な発生間隔を監視し，問題が起きつつあるかどうかを発見する．すなわち，クラスタの要素の時系列変化を調べ，過去と比べてここ最近にメー

ルが来ている頻度が高い場合は，問題が起きつつあると判断する．方法論としては，バースト検出アルゴリズムを用いる．時系列データにおいてデータ間の時間間隔が平常時よりも密になっている状態をバーストと呼び，バーストを発見する手法がバースト検出である．なおクラスタリングとバースト検出を合わせて統計的要約と本研究では言うことにする．

情報可視化

情報可視化とは様々な情報を分かりやすく表示することでユーザの理解を促す技術の総称である．ここで分かりやすいとは直感的に理解できる表示であること，簡便なインタラクションを備えていることを言う．表示においてもこれらの点を考慮する必要がある．これらの事を念頭に置き可視化手法を検討する必要がある．

まずは表示方法について述べる．１つのアイコンで複数の属性情報を表現した場合，それらの情報をテキストで補足的に表した場合に比べて情報を認識する速度が上昇したという研究がある[12]．そこで本研究でもアイコンに複数の属性情報をもたせることにする．具体的にはカテゴリ情報を色，事象の規模を推定しやすいようにクラスタサイズに応じてアイコンを大きく表示する．

次にインタラクションについて説明する．データベースに蓄えられた情報を取り出した時，情報の数やオペレータの意図に応じて地図上での提示方法を自由に変えられるようにする．具体的には図１のように可視化表示に３つの状態を備える．状態１はクラスタリングを行わず，１つの情報につき１つのアイコンを対応させて表示させる方法，状態２はクラスタリングにより良く似た情報を１つにまとめて表示する方法，状態３は各クラスタに対してバースト検出を行い，現在起こっている問題だけを表示する方法である．これら３つのモードを状況に応じてオペレータがインタラクティブに操作する事ができる．状態１から２は周辺状況の概要を把握したい場合，逆に２から１は細部を確認したい場合がユーザがインタラクションを起こす上でのモチベーションとなる．また状態２から３は現在起こっている問題が知りたい場合，３から２は過去の問題も含めて知りたい場合がモチベーションとなる．

4. システムの実装

本章では前章で説明したシステム処理の実際について述べる．

4.1 システムの概要

本研究で設計するシステムの概要は図２のように表される．まず一般市民が周辺状況をＧＰＳ情報を添付してメールで送り，メールの本文，送信時間，位置情報（緯度・経度）をデータベースに蓄える．送信時間はメールのヘッダ情報から得られ，位置情報は GPS 情報

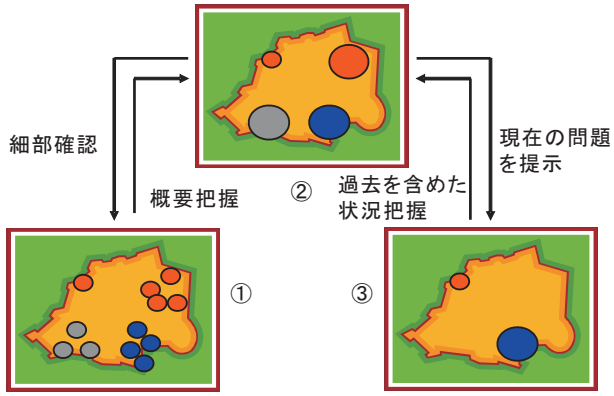

図 1 可視化表示時のインタラクション

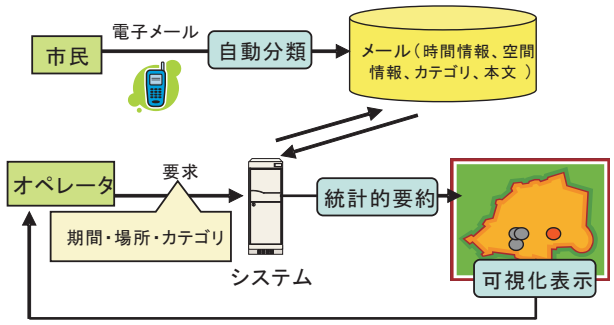

図 2 システムの概要

から取得する．またあらかじめメールがどのような内容のものかカテゴリを設定しておき，メールのテキストを自動分類することにより，メールがカテゴリに属するか否かを決定する．そしてその結果についてもデータベースに格納する．

次にデータベースに蓄積された情報について，まずオペレータが期間・場所・カテゴリを指定して，オペレータが欲しい情報のみを取り出す．すると取り出された情報がシステムにより可視化されてオペレータに提示される．前章で説明したとおりオペレータは情報の地図上への提示方法については自由に変更する事ができる．

4.2 メール本文の自動分類

メール文の自動分類を行う上で必要な分類器を，本研究では以下のようにして生成する．

Step1:業務内容に合うようシステムの設計者がカテゴリを決定

Step2:学習用データを用意し，正解データを人手で付与

Step3:学習用データに対して形態素解析と属性選択を行い特徴ベクトルを生成し，機械学習を行う事で分類器を生成.

Step2 ではカテゴリに属するか否かが曖昧なメール文も多く存在するので，複数で正解データを付け，人による正解データの違いを吸収できると良い．Step3 における機械学習では SVM[7] を用いて分類器を生成する．SVM は高次元に強く，汎化性が高い事が知られておりテキストの自動分類の研究では頻繁に使われている．形態素解析のツールとしては Sen[8] を使用し，SVM はツールとして libsvm[9] を使用する．

属性選択の問題

テキスト自動分類において学習器を生成する時，どの単語を用いて学習器を生成すれば高い分類精度が得られるのかを考える属性選択の問題がある．学習を行う時，各カテゴリに関連する単語をそれぞれ各カテゴリの分類に用いる属性とすれば良い．しかしあまり属性数が多いと過学習が起こりやすくなりかえって分類精度が悪化してしまう．そこで各カテゴリに関連する単語を関連する度合いが大きい順に並べ，上からどれだけ属性として選択すれば（すなわち属整数をいくつにすれば）最も良い分類精度が得られるのかを調べ，属性を決定する．具体的には以下の方法で最適な属性を決定する．

最初に最適な属整数を求める．まず学習用データを属性数決定用の学習用データと属性数決定用のテスト用データに分割する．次に属性数決定用の学習用データのメール文を形態素解析することにより単語に分け，単語そのものが何らかの意味を表す内容語（本研究では名詞，動詞，形容詞）を属性の候補として選択する．次に選択した属性の候補と各カテゴリとの相互情報量を求め，カテゴリ別に相互情報量が高い順に単語を並べる．単語 t とカテゴリ c の間の相互情報量 (MI) は式 (1) で表される．

$$MI(t;c) = P(t,c)log\frac{P(t,c)}{P(t)P(c)} \quad (1)$$

ここで $P(t)$，$P(c)$，$P(t,c)$ はそれぞれ全メール文中での単語 t を含む文の割合，カテゴリ c に属する文の割合，単語 t を含みかつカテゴリ c に属する文の割合である．相互情報量は t の出現頻度がある１つのカテゴリ c とその他のカテゴリの間で偏りがある時に大きな値となる．すなわちあるカテゴリに対するある単語の相互情報量が高い事はそれだけその単語がそのカテゴリにおいて関連する単語であることを意味する．次に相互情報量が高い順に並べられた単語を上からいくつ属性として選択すれば，すなわち属性数をいくつにすれば最も分類精度が良くなるのかを調べる．手法としては属性数決定用の学習用データから属性数を変化させつつ学習器を生成し，属性数決定用のテスト用データを用いて調べる．分類精度の基準としては F 値を用いる．最も F 値が高くなるときの属性数を最適な属性数として決定する．

最適な属性数が決まると次は学習用データ全て（属性数決定用の学習用データと属性数決定用のテスト用データ）を用い，単語を相互情報量が大きい順に並べ，最適な属性数の数だけ上から単語を選択し属性を決定する．

4.3 統計的要約

クラスタリング

良く似た情報を１つにまとめるためにクラスタリングを行う．良く似た情報とは空間的，時間的，内容的に近い場合が多い．よって基本的にはこの３つを要素としてクラスタリングをして情報をまとめてやれば良い．しかしより正確に情報をクラスタリングするため幾らか工夫する必要がある．

まず内容の類似度を測るにあたってテキストで使われる単語を属性として情報間の類似度を測れば良いと考えるかもしれない．しかし携帯から送られる文章はニュース記事や Web サイトのページと比べて単語数が少なく，単純に２つの情報間のテキストを比べるだけでは同じ単語が使われている確率が低く十分な精度を出す事ができない．そこでオペレータの重視する内容を表すカテゴリを利用する．カテゴリは内容を大まかに表したものなので，２つの情報間でカテゴリがどの程度重複しているのかを調べる事で内容の類似度を測る事ができる．

次に空間と時間の扱いについて考える．まず同じ時間・カテゴリのイベントでも場所が異なると全く別のイベントになる事もある．例えば同じ水害カテゴリに属する情報であっても異なる場所で発生した水害は，それを引き起こした原因は異なるかもしれない．よって情報間の空間的距離はクラスタリングを行う上で基準のひとつとすべきである．次に時間についても同じ場所，カテゴリであっても異なる時間に来たメールは別のイベントになる可能性はある．しかし同じ場所で異なるイベントが何度も起こり, そのそれぞれが重要となるケースは少ないように思われる. それよりはイベントの時間的な変化の方が重要と考える．このため時間情報はクラスタリングでは使わず，バースト検出で利用する事とする．以上により空間情報とカテゴリ情報を用いクラスタリングを行う．

クラスタリングには Newman[10] のクラスタリング手法を適用する．この手法は凝集法の一種である．凝集法ではクラスタの併合をどの段階で止めるかという問題が発生するが，Newman はモジュール性という概念を用いてこの問題に関する有効な解決策を提示している．モジュール性はクラスタリングの分割の良さを表す指標であり，主にネットワークから密につながったノードのまとまりを発見する，すなわちコミュニティを発見する時に用いられる．またノード間のエッジの重みを考慮する事によってよりきめ細かいクラスタリングを実現する事ができる．本研究ではメール１つを１ノードとし，メールの送信位置に基づく距離とメール本文の内容の類似度からノード間のエッジの重みを算出しクラスタリングを行う．

定式化は以下のように行う．モジュール性を測るモジュール関数を Q とする．ノードの集合を V，エッジの集合を E，そして重み付き隣接行列を $W = [w_{i,j}]_{n \times n}$($n$ はノードの数からなる無効グラフ $G(V, E, W)$ を，k 個のコミュニティ $(V1, V2, \ldots, V_k)$ にクラスタリングした際のモジュール関数は次のように定義される．

$$Q = \sum_{c=1}^{k} \left[\frac{L(V_c, V_c)}{L(V, V)} - \left(\frac{L(V_c, V)}{L(V, V)} \right)^2 \right] \tag{2}$$

ただし，$L(V^{'}, V^{"}) = \sum_{i \in V^{'}, j \in V^{"}} w(i, j)$ でありi,j はメールの番号である．もし，すべてのコミュニティが完全にランダムに生成された場合 $Q = 0$ となる．より密に繋がったコミュニティに分割できているほど，Q は 1 に近づく．まず最初に各ノードを１つのクラスタと見なし，各ステップにおいて２つのクラスタを併合していく．併合の際は，Q を最も増加させるクラスタのペアを選択する．クラスタ V_i とクラスタ V_j を併合する時の Q の増加分 ΔQ は以下のように与えられる．

$$\Delta Q = 2 \left(\frac{L(V_i, V_j)}{L(V, V)} - \frac{L(V_i, V)}{L(V, V)} \times \frac{L(V_j, V)}{L(V, V)} \right) \tag{3}$$

Q の最大値を与えるステップを最終的なクラスタリング結果とする．

またエッジの重みについては以下のようにして算出した．エッジの重みには２つのメールの送信位置間の距離から算出する重み w_d と２つのメール本文の内容の類似度から算出する重み w_c の積

$$w = w_d \times w_c \tag{4}$$

を用いる．ユーザがデータベースにアクセスし，取り出された全メールの中で，i 番目に取り出されたメールと j 番目に取り出されたメールとの距離を $d(i, j)$ とおき，そのうち最も近接したメール間の距離を 1 とおくと $w_d(i, j) = l/d(i, j)$ となる．一方メール本文の内容の類似度による重みは，カテゴリの重複度合いをコサイン距離で求める事によって算出する．コサイン距離は異なる２つのメールから得られるカテゴリの特徴ベクトルをそれぞれ X, Y とおくと（例えば１から６まで６つのカテゴリを設定していたとして，１つのメールがカテゴリ１とカテゴリ５に属するとその特徴ベクトルは $(1, 0, 0, 0, 1, 0)$ で表される.）コサイン距離による重みは $w_c(i, j) = \frac{X \cdot Y}{\sqrt{|X||Y|}}$ と表される．

バースト検出

現在正に起こっている問題を検知するため，我々は Kleinberg[11] のバースト検出手法を用いる．この手法は document stream におけるバーストを検出する手法である．document stream とは時系列の文書の流れであり，例えば電子メールやニュース記事，掲示板などがこれに相当する．document stream において，ドキュメントが送られてくる時間が平常状態よりも密になっている状態をバースト状態と呼ぶ．連続してバースト状態にあるドキュメント群からは，問題が顕在化，あるいは深刻化している可能性が読み取れ，バーストを抽出する事でいち早く現在起こっている問題を知る事ができる．Kleinberg の提案する手法では，状態遷移コストを含めたモデルや含めないモデル，バーストをその度合いにより多段階に分けたモデルなど様々な定式化手法が述べられている．我々のシステムでは状態遷移コストを含め，バースト状態と平常状態の２状態からなるモデルで定式化して最尤状態列を求め，更に新たにバーストの大きさを測る基準であるバースト度を定義する．そしてバースト度が最大となった時のメールを検出し，その時刻を新たな問題の発生時刻として定義する．システムを使用する側は，例えば「３０分以内に新たに発生した問題について知りたい」という要求をシステムに入力したとすると，システムは問題発生時刻が過去３０分以内のものを地図上に表示する．

定式化は以下のようにして行う．まず２状態からなるオートマトン A を定義し平常状態を q_0，バースト状態を q_1 とおき，時間 T の間に $n+1$ 個のメールが到着するとする．ただし T は最初のメールが到着してから最後のメールが到着するまでの時間とする．もしメールが等間隔で到着した場合，その間隔を $\hat{g}$ とおくと $\hat{g} = T/n$ となる．またメールがランダムに到着すると考えるとあるメール i が到着してから次のメール $i+1$ が到着するまでの間隔は指数分布に従う事になる．この時間間隔を x とおき，また $\alpha_0 = n/T$ とおくと間隔 x で次のメールが到着する確率は $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ となる．これを平常状態での確率密度関数とする．バースト状態の時は平常状態より短い時間間隔でメールが送られてくるため

$\alpha_1 = s\alpha_0 (s > 1)$ として，間隔 x で次のメールが到着する確率は $f_1(x) = \alpha_1 e^{-\alpha_1 x}$ であるとする．また $n+1$ 個のメールが到着する一連の間隔を $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)$ とおき $x_i > 0$ とする．そして各メール間隔での状態を $\mathbf{q} = (q_{i_1}, q_{i_2}, \ldots, q_{i_n})$ とする．例えば k 番目のメールが到着してから $k+1$ 番目のメールが到着するまでの状態は q_{i_k} とおける．時間間隔が $\mathbf{x}$ の時に状態列が $\mathbf{q}$ になる確率が最も高くなるような $\mathbf{q}$ が求める最尤状態列である．すなわち

$$Pr[\mathbf{q}|\mathbf{x}] = \prod_{t=1}^{n} f_{q_t}(x_t) \tag{5}$$

が最も大きい値を選べば良い．これは次の値を最小にすることと等価である．

$$-\ln Pr[\mathbf{q}|\mathbf{x}] = \sum_{t=1}^{n} -\ln f_{i_t}(x_t) \tag{6}$$

これが最小となるような状態列 $\mathbf{q}$ を求めれば良いのだが，このように判定してしまうと平常状態とバースト状態が簡単に移り変わってしまう．すなわち細かなバーストが数多く発生する事になり，ある程度まとまったバーストを検出したいような場合に不利である．そこで状態遷移コスト $\tau(l, j)$ を導入する．l, j は状態を表す文字であり，0(平常状態）または 1(バースト状態）を取る．そして平常状態からバースト状態に遷移する時とバースト状態から平常状態へ遷移する時に γ を係数としてコスト $\gamma \ln n (\gamma > 0)$ が発生するものとする．すなわち

$$\tau(l, j) = \begin{cases} \gamma \ln n (l \neq j) \\ 0 (l = j) \end{cases} \tag{7}$$

である．よって式 (6) と (7) を合わせて次の最小コストを求めれば良い

$$C[\mathbf{q}|\mathbf{x}] = \left(\sum_{t=1}^{n} -\ln f_{i_t}(x_t) \right) + \left(\sum_{t=0}^{n} \tau(i_t, i_{t+1}) \right) \tag{8}$$

具体的には以下のような手順で計算を行う。$C_{i_t}(t)$ を状態が q_{i_t} で終了する入力 $\mathbf{x} = (x_1, x_2, \ldots, x_t)$ に対する状態列の最小コストであるとすると

(1) 初期状態 $t = 0$ において $C_0(t) = 0$，$C_1(t) = \infty$ とする
(2) t = t + 1
(3) $C_{i_t}(t) (i_t = 0, 1)$ を計算する．

$$C_{i_t}(t) = -\ln f_{i_t}(x_t) + \min_{i_{t-1}} (C_{i_{t-1}}(t-1) + \tau(i_{t-1}, i_t)) \tag{9}$$

(4) 全てのメールについて (2)，(3) を繰り返す．また前状態が平常状態であったか，バースト状態であったかを保持しておく．
(5) コストが最小である状態列を選択する．まず $C_{i_n}(t) (i_n = 0, 1)$ で値が小さい方 (この最小コストを $minC(\mathbf{x})$ とする) を選択し，その状態の前状態を繰り返し辿る事で状態列が選択できる．

またある k 番目のメールがバースト状態ではなく平常状態にあったと仮定し全体のコストを再計算した時 (このコストを $minCk(\mathbf{x})$ とおく) に生じたコストの増大をバースト度 $B(k)$ と定義する．すなわち

$$B(k) = minCk(\mathbf{x}) - minC(\mathbf{x}) \tag{10}$$

ここでバースト度が最も高い時の k を求め，その時刻を新しい問題が起こった時刻として扱う．

4.4 Web ベースでのシステム実装

本研究では，導入や実装が簡単に行え，保守も容易であるという点から Web ベースでシステムを実装する．地図は Google マップを用い，Google Maps API を用いて実装を行う．またよりスムーズな情報提示のために Ajax を用いる．Ajax を用いることでデータベースから必要な情報を取り出して表示する際にページ遷移を伴わないで表示させることができる．図 3 は情報の表示例である．まだ開発中のものだが，画面左側では Google マップ上でアイコンが表示されており，アイコンをクリックする事により詳細情報が得られる．画面右側ではアイコンの持つ情報が一覧にして表示されているのがわかる．

5. 実　　験

本章ではシステムの有効性を調査する．評価には大量の時空間情報を伴うテキストデータが必要である．ところが現状ではそれほど多くの人が携帯電話で周辺状況について送信しているわけではないため，大量の情報を取得する事は不可能である．そこで万博記念公園において被験者に携帯電話で周辺状況について送信してもらいデータを収集することにした．このデータを元にして

(1) 我々が公園の管理を行うことを想定して自動分類のためのカテゴリを決定し，十分な分類精度が得られるかを調べる．
(2) 統計的要約においてクラスタリング，バースト検出が上手く機能するかどうか調べる．

図 3　表示画面の例

(3)　システム全体を実装し，システムが状況の把握に有効であるかを検証する．

を行う．本論文ではデータの収集と自動分類まで行ったので，そこまでの結果をまとめる．

5.1　データの収集

万博記念公園にて周囲の状況をメールで GPS 情報を添付して被験者に送信してもらう．GPS 機能がついた携帯電話を所持していない被験者は紙の地図でメールを送信した位置を書き記してもらうこととした．送るメールの内容はユーザが役立つ情報なら何でも良く，特に指示しないこととした．ただし被験者が何を送ってよいかわからないと困るので例文だけは用意しておいた．日時は 2008 年の 11 月 16 日の 9 時半から 16 時半までの 7 時間，被験者数は 120 人であった．またこの日は無料デーで，ラジオの放送やフリーマーケットなど多くのイベントが行われていた．

結果として合計で 2034 件のメールを得た．このデータに対して自動分類，統計的要約が有効に機能するかについて調べていく．

5.2　電子メールの自動分類

本実験では公園の管理者の視点に立ち次のようなカテゴリを設定した．

１．施設　２．遊び・スポーツ・体験　３．鑑賞

４．食事　５．人々の状況　６．イベント

施設はトイレや地面の状態，交通手段，売店の状況など場所や移動手段に関する情報を，遊び・スポーツ・体験は自然体験学習やアスレチックなど人が主体的に行うことに関する情報を，鑑賞はショーや展覧会など受動的に見ることに関する情報を，食事はレストランや屋台などの食べ物に関する情報を，人々の状況は周りの人の数や行動に関する情報を，イベントは万博記念公園で開催されている催し物に関する情報をそれぞれ表す．電子メールの自動分類ではそれぞれのカテゴリについて属す，属さないの２値で分類する．2033 個のデータをランダムに１０等分して，半分の５セットを学習用データとして用い，残りの５セットをテスト用データとして用いた．また正解データは一般性を持たせるために，万博公園でのメール送信の被験者とは異なる新たな５人の被験者に各データが各カテゴリに属するか否かを判断してもらい，３人以上が属すると判断すれば属するとした．

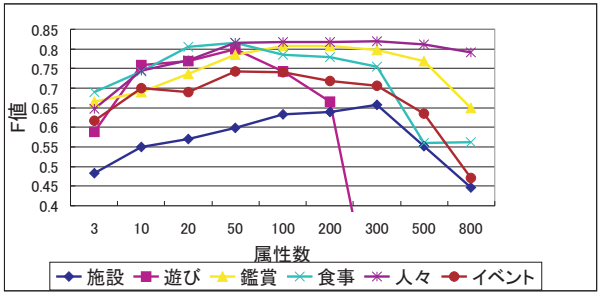

図 4　属性数による F 値の変化

次に最適な属性数を求めるため，学習用データのうち４セットを属性数決定用の学習用データとして用い，残りの１セットを属性数決定用のテスト用データとして用いた．そして属性数決定用の学習用データから 4.2 で説明した手法で各カテゴリに関連する単語を相互情報量の順に並べ，属性数を各カテゴリにおいて 3，10，20，50，100，200，300，500，800 と変化させて属性を選択し，それを元に学習器を生成し，5-クロスバリデーションにより F 値を求めたところ図 4 のようになった．

図より最適な属性数はそれぞれ施設が 300，遊び・スポーツ・体験が 50，鑑賞が 100，食事が 50，人々の状況が 300，イベントが 50 となり，かなりカテゴリによって属性数が異なる結果となった．次に全ての学習用データを用いてカテゴリ別に単語の相互情報量を求め，相互情報量が高い順に最適な属性数の分だけ属性を選択し，この属性を用いて学習器を生成した．表 1 は選択した属性の例である．色のついている部分が選択した属性である．

この属性を用いて生成した学習器とテスト用データを用いて自動分類結果を求めたところ表 2 のような結果なった．確実な情報を求めるのであれば F 値は十分とは言えないが，全体として何が起こっているのかを概観するという用途であれば十分な結果であると思われる．「人々」カテゴリの F 値が高くなった理由として２つ考えられる．１つ目は「人々」のカテ

表 1　相互情報量によって選択した単語

	施設	遊び	鑑賞	食べ物	人々	イベント
1番	休憩	ボート	紅葉	ラーメン	人	ラジオ
20番	民族	滑り台	展示	粕汁	たち	北海道
50番	森	バ	チューリップ	匂い	天国	民族
100番	迷子	こぎ	非常	食	動き出す	はじまる
300番	周囲	冷たい	必見	だす	付き	帰路
500番	っぽい	なかなか	着く	駅	終える	怪しげ

表 2　自動分類の再現率・精度・F 値

	施設	遊び	鑑賞	食事	人々	イベント	平均
再現率	0.618	0.758	0.724	0.723	0.877	0.736	0.739
精度	0.694	0.688	0.740	0.752	0.792	0.766	0.739
F値	0.656	0.723	0.732	0.737	0.834	0.751	0.739

ゴリに属するメール文が多く十分な学習が行えたためでだと考えられる．二つ目は正解データをつける時に「人々」カテゴリに属しているか否かの判断が行いやすく，人によって差異が生じず安定した正解データが得られたためだと考えられる．例えば「遊んでいる人がいます」や「写真を撮っている人が多いです」などのメール文が多く，人々のカテゴリに属するか否かを判定しやすいものが多くあった．一方「施設」カテゴリはその逆で，学習量が少ない上に，正解データの判定が曖昧なものが多かったため F 値が低くなったのだと考えられる．判定が曖昧であった例としては,「太陽の塔の前にいます」や「列車が通ります」などが挙げられる．太陽の塔や列車が通ることを施設と考えるかは人によって意見が分かれた．

6. まとめと今後の課題

時空間情報を伴う大量のテキストデータから業務上重要な情報だけを抽出し，状況を把握しやすいよう要約・可視化するシステムを設計した．またその評価として今回は電子メールの自動分類を行った．今後の課題としては統計的要約でのクラスタリング，バースト検出の精度評価を行い，被験者実験によりシステム全体の評価を被験者実験を通して実験していきたい．

参 考 文 献

1) twitter:http://twitter.com/
2) 上松大輝，沼晃介，徳永徹郎, 大向一輝, 武田英明：場 log: Weblog 環境における位置情報利用の提案，第 6 回人工知能学会セマンティック Web とオントロジー研究会, (2004).
3) Google マップ：http://maps.google.co.jp/
4) YAHOO!地図：http://map.yahoo.co.jp/
5) 倉島健，手塚太郎，田中克己：街 Blog からの体験抽出とその空間提示手法の提案．電子情報通信学会技術研究報告，Vol.105, No.171, pp.35–40 (2005).
6) 宮森恒，水口充，河合由起子，是津耕司，木俵豊：雰囲気メタファによる町の偏在情報の集約・提示システムの検討，第 1 4 回 Web インテリジェンスとインタラクション研究会，pp.101–102 (2008).
7) Cortes, C. and Vapnik, V.:Support Vector Networks, Machine Learning, Vol.20, No.3, pp.273-297 (1995).
8) sen:http://www.mlab.im.dendai.ac.jp/yamada/ir/MorphologicalAnalyzer/Sen.html
9) libsvm:http://www.csie.ntu.edu.tw/cjlin/libsvm/
10) M. E. J. Newman: Detecting community structure in networks，Eur. Phys. J. B, Vol.38, No.2, pp.321-330 (2004).
11) Jon Kleinberg: Bursty and Hierarchical Structure in Streams, In Proc. the 8th ACM SIGKDD, pp.91–101 (2002).
12) Spence, R., Parr,M.: Cognitive Assessment of Alternatives,Interacting with Computers,3,3,pp.270–282 (1991).
13) Google maps API: http://code.google.com/intl/ja/apis/maps/