2005－EIP－27（4）
2005／3／19

Web 文書タイプ自動分類手法の比較評価と適用

井筒 清史† 　横澤 誠‡ 　篠原 健‡

†京都大学大学院 情報学研究科 社会情報学専攻
‡(株)野村総合研究所

Web の高度な利活用のニーズとして，社会的な出来事に関して Web 空間上での Web ページ数の反応の把握がある．このニーズに応える手段として，Web ページをインデキシングする技術である Web の自動分類手法を用いることにした. 本稿では, Web の自動分類器を作成し，社会的な出来事を把握するために利用し，その技術の有効な利用シーンを提示する．Web の自動分類器を作成するために，判別分析，Rule base，Naïve Bayes の代表的な 3 つの手法の応用領域での性質について調べた．Web の自動分類手法の有効な利用方法を提示するため，ここでは特に Rule Base 分類器によるスマトラ島沖地震とノートパソコンの新製品販売・発表に関する時系列分析の実験を行った．各出来事が起こる前後で，予め定めたカテゴリーごとに Web ページの数に特徴的な反応があることが観測された．

Comparative Evaluation and Applications of Automatic Web-based Document Classification Methods

Kiyoshi Izutsu† 　Makoto Yokozawa‡ 　and 　Takeshi Shinohara‡

†Department of Social Informatics, Kyoto University
‡Nomura Research Institute, Ltd.

Numbers of web pages about social incidents are expected to be a good index of how the incident is considered in web communities. Possibilities were shown in this study that automated web page classification method can help those numerical analysis of web space. Discriminant Analysis method, Rule-Based method and Naïve Bayes method were compared by their classification accuracy, developers' cost, users' cost, processing time and scalability. Practical experiments about the Sumatran Earthquake/Tsunami and a new Note PC Product show communities response differently around the time of the incidents.

1. はじめに

近年，Web ページの数が爆発的なスピードで増加している．そのため，必要な Web ページを容易に利活用することが困難になってしまった．この現状に対応し，Web 上の情報をユーザの目的に合った利活用が可能となる手段を提供するものとして様々な技術が開発されてきている．

このような技術のひとつとして, Web の自動分類が注目を浴びている．この技術は，特定の URL 集合が与えられたときにユーザの利用目的に合わせて予め作成された文書タイプ(カテゴリー)に Web ページを振り分け，インデキシング(Web ページにカテゴリー情報を付与すること)するものである．

Web の自動分類技術は, 既存の技術と連携すること

によって，その有効性を発揮するものと考えられる．サーチエンジンの技術と Web の自動分類手法とが連携したサービスとして，Vivisimo[1]，Teoma[2]，Grokker[3]などが登場している．これはサーチエンジンの検索結果を任意のカテゴリーに分類して提示する新しい情報検索の提案として注目を浴びている．このように Web 自動分類の技術を他の技術に応用することによって，ユーザは利用目的に即した高度な Web の利活用へつなげることが可能になると考えられる．

本研究では，社会的な出来事に関して，Web 空間上でのカテゴリーごとの反応を把握するという Web の自動分類の適用分野を提示することを目的とする．Web の自動分類器を作成するために，判別分析分類器，Rule Base 分類器，Naïve Bayes 分類器の 3 つの手法の性質の評価を行なった．ここでは特に Rule Base 分類器を選択し，スマトラ島沖地震とノートパソコンの新製品発表・販売の出来事に関して，Web 空間上での Web ページの数の反応を観測することにした．

2. 関連研究

Web の自動分類に関する関連研究を任意のカテゴリーに分類するものと既定のカテゴリーに分類するものに分けて考える．任意のカテゴリーに分類するものとしては，Vivisimo[1]，Teoma[2]，Grokker[3]などが実際のサービスとして登場している．また URL のツリー構造やレイアウト情報で Web の分類を行うもの[4]もある．既定のカテゴリーに分類するものとしては，Web の構造的な特徴をもとに予め決められた文書タイプに分類するもの[5](本研究の Rule Base 分類器はこの研究の成果を利用している)や，リンク構造とテキスト処理を組み合わせて高精度に Web ページを分類するもの[6]が登場している．本研究は，既定のカテゴリーへ Web ページを分類するものに属していて,人間社会の出来事に関して Web 空間上での Web ページの反応を把握することが可能であることを示す．

3. Web の自動分類手法

本章では，Web を自動分類する際の分類先のカテゴリーと自動分類手法として用いた判別分析分類器，Rule Base 分類器，Naïve Bayes 分類器の解説を行う．

3.1. カテゴリー

社会的な出来事が発生したときに，Web ページの数の反応が人間にとって意味のある分類になるように，カテゴリーを選択する．ここでは Web ページの持っている機能的側面に注目した．例えば，ニュースサイトは，メディアが企業や個人に対して情報を伝達する機能を持っているなどである．Google でノートパソコン(VAIO，ThinkPad，BIBLO，LaVie，dynabook) [1]と航空会社(ANA，JAL) [2]を検索し，延べ 2000 ページを調査して， 以下の 6 つのカテゴリーへ分類したいユーザがいる場合を想定して実験をすすめることにした．

- **enterprise**：企業が製品やサービスを案内している Web ページ
- **news**：報道媒体が情報提供を行っているページ
- **shop**：製品やサービスを提供，販売している Web ページ
- **diary**：主に個人が日記により情報を発信している Web ページ
- **bbs**：特定のテーマで個人同士，自由に意見を書き込んでいるコミュニティサイトの Web ページ
- **others**：日記系以外の個人，検索サイト，用語辞典などを含む上記のカテゴリー以外のページ

3.2. 判別分析分類器

ユーザはある Web ページを見たとき，そのページが「ショッピングサイト」であるとか「掲示板サイト」であるといったことを一瞥しただけで理解することができる．例えば，ショッピングサイトでは，商品を購入するために必要なフォーム，価格を表す表示などのレイアウトが現れる場合が多い．ここで，レイアウトとは Web ページの画像や動画，入力フォームなどの見かけ上の配置のことを言う．

---

[1]VAIO はソニー, ThinkPad は IBM, BIBLO は富士通, LaVie は NEC，dynabook は東芝の商標である．
[2]ANA は全日空，JAL は日本航空の商標である．

表 3-2 分類精度が最も良かった Web 構造情報一覧

Web 構造情報	特徴量名称	説明
テキスト	掲示板関連の文字列	特定のタグで囲まれた掲示板,bbs などの文字列の数
	商品価格を表す文字列	特定のタグで囲まれた￥1,000 や 100 円などの数
	ショッピングカート	ショッピングカートを表現する文字列や画像の数
フォーム	input の数	入力フィールドを作る．<input>の数
情報量	kbyte 当たりのリンク数	リンク数/kbyte 数
リンク	自サイトリンク数の割合	自サイトリンク数/総リンク数
	trackback の数	blog サイトなどに見られる trackback の数
画像	広告系 img の数	横 728，縦 90 の画像の数など

上述してきたことを踏まえ，判別分析分類器による Web の自動分類は，「Web ページのレイアウト情報を統計手法にかけることにより, 大量の Web ページを短い時間内に実用的な精度で分類できるかもしれない」という仮説を検証するために行った． HTML ソースのレイアウト情報から特徴量を抽出し，説明変数として判別分析にかけることによって，Web ページをカテゴリーへの振り分けることにした．以下，特徴量を構成する Web 構造情報の定義の説明を行う．

3.2.1. Web 構造情報

本稿では Web ページのレイアウトを表現している HTML ソースの情報を Web 構造情報として定義する．

ノートパソコンに関する Google の検索結果をもとに，延べ 1500 ページを調査し，Web の分類を行うために利用可能な Web 構造情報を選択した．4 章で解説する評価実験の結果, 最も分類精度よかった Web 構造情報を表 3-2 に示す．

3.3. Rule Base 分類器

Rule Base 分類器による Web の自動分類は，「各カテゴリーには，そのカテゴリーを特徴付ける機能語というものが存在して，それらを抽出することによって Web の自動分類にいかせるであろう」という仮説を検証する．機能語とは，各カテゴリー特有の機能を端的に表している文字列である．機能語は，HTML タグ内の文字列や HTML タグに囲まれた文字列をパースして特徴量として抽出される．例えば，企業サイトは，リリースしている製品に関する製品情報という文字列をアンカーテキストとして配置する場合が多い．

表 3-3 機能語の特徴量一覧

カテゴリー	機能語	出現場所と得点 (特徴量)
enterprise	製品情報	<title>：5，<a href>：3
	Inc. ltd. copylight	タグによらず：2
news	編集部	<a href>：3
	コラム	<a href>，<b>：2
shop	ショッピングモール	<title>：4
	￥1,000(価格,円も含む)など	<font>, <b>,, ：1
diary	日記	<title>：4，<a href>：2
	trackback	<a href ” ”>：5
bbs	掲示板	<a href>,<a href ” ”>：5, <img alt=” ”>，<font>：3
	ファンサイト	タグによらない：3
others	リンク集	<title> ： 6,<a href> , <font>,,:4
	ショップガイド	<a href>：4

このように Rule Base 分類器による Web の自動分類では，ユーザが Web ページをどのようなページか判断するのと同じように，カテゴリーに典型的な文字

列(一部画像などを含む)を特徴量として抽出し，総合的に判断して分類する．

6 つのカテゴリー(enterprise，news，shop，diary，bbs，others)の特徴量の例を表 3-3 に示す．表 3-3 に示したように，カテゴリーごとに決められた HTML タグの内外に特定の文字列が出現した場合，各カテゴリーに用意した変数の得点を加算していく．HTML ソースから特徴量の抽出が終わったとき，すべてのカテゴリーの中で最も得点の高かったカテゴリーへ Web ページを分類することにした．

3.4. Naïve Bayes 分類器

Naïve Bayes 分類器を用いた Web の自動分類は，HTML ソースに含まれる単語によって特定のカテゴリーに HTML ソースを分類する確率を計算する手法である．カテゴリーは単語とそれぞれの単語の頻度のリストで構成されている．いくつかのカテゴリーをあわせたものをコーパスと呼び，HTML ソースを分類するカテゴリー，HTML ソースに含まれる個々の単語が特定のカテゴリーに含まれる確率，HTML ソースがカテゴリーに含まれる確率を決定する．C_1～C_n のカテゴリーがあり，W_1 から W_m までの m 個の単語があるとする．ここで，特定の HTML ソース H が含まれるべきカテゴリーを計算することになる．まず，それぞれのカテゴリーCi について，P(Ci|H)を計算する．これは，Bayes の定理を用いると以下のように計算される．

$$P(Ci \mid H) = \frac{P(H \mid Ci) \cdot P(Ci)}{P(H)}$$

ここで，P(Ci|H)は HTML ソース H がカテゴリー Ci に含まれる確率である．これは，H に含まれる単語のうち，カテゴリーCi に現れる単語から計算されている．P(H|Ci)はカテゴリーCi について，H に含まれる単語がカテゴリーの中に現れる確率である．P(Ci)は与えられたカテゴリーが選ばれる確率で，カテゴリーCi に含まれるいずれかの HTML ソースが選ばれる確率でもある，P(H)は特定の HTML ソースが現れる確率を表す．

H をどのカテゴリーに分類するか計算するためには，P(Ci|H)をそれぞれのカテゴリーについて計算し，最大のものを探す．どの計算においても，P(H)を使用するので，これを無視して，

$$P(Ci \mid H) = P(H \mid Ci) \cdot P(Ci)$$

を計算することにする．

まず，H を H に含まれる単語リストに分割して，これらを H1~Ho と呼ぶ．P(H|Ci)を計算するためにはそれぞれの単語の確率の積を計算する．これはそれぞれの単語が Ci に含まれる確率である．Naïve Bayes 分類器では，単語が出現する確率は他の単語とは関係なく独立であると仮定している．P(Ci)は Ci に含まれる全ての単語の数を，すべてのカテゴリーの単語を足し合わせたもので割ることによって計算される．最終的に，P(Ci|H)はそれぞれのカテゴリーについて

$$P(Ci \mid H) = P(H1 \mid Ci) \cdot P(H2 \mid Ci) \cdot P(H3 \mid Ci) \cdots P(Ho \mid Ci) \cdot P(Ci)$$

として計算され，そのうちから最大のものを選ぶことになる．[7]

4. Web の自動分類器の評価実験

本章では Web の自動分類手法の性質を比較・検証するために行った評価実験について解説していく．第 3 章で解説した判別分析分類器，Rule Base 分類器，Naïve Bayes 分類器の 3 つの手法を用いて Web ページを 6 つのカテゴリーへ分類する．はじめに評価実験の目的と対象について述べる．次に評価実験の手順，結果の順に述べていく．

4.1. 評価実験の目的と対象

分類性能を比較・検証するために行った評価実験について解説する．評価すべき 5 つの項目を以下に述べる．

- 各カテゴリーへの分類精度
- 分類精度が得られるまでの開発者の負担
- ユーザの負担
- システムに対する負荷(計算時間)
- スケーラビリティ

評価実験によって，この 5 つの項目を検証していく．

表 4-1　テストコレクション「ノートパソコン」のページ数

カテゴリー	enterprise	news	shop	diary	bbs	others	Total
ページ数	140	100	50	40	20	35	385

表 4-2 評価実験の結果まとめ

分類手法	判別分析分類器	Rule Base 分類器	Naïve Bayes 分類器
分類精度	正解率 83%	正解率 89%	正解率 78%
開発者の負担	各カテゴリーを特徴付ける細かなレイアウトを見つけていく必要がある(11 時間)	単独のカテゴリーを特徴付ける Rule 作成が難しい(約 250 の Rule を作成，12 時間半)	特になし
ユーザの負担	判別関数を作成するデータを用意する必要がある(約 1 時間)	特になく	精度を向上させるには数多くのユーザのフィードバックが必要(144 回)
システムの負担	高速(52 秒)	高速(88 秒)	低速(693 秒)
スケーラビリティ	レイアウト情報の変動により，分類精度が極端に低下ため「低い」	機能語の表現の変化により，分類精度が低下する．Rule の追加で対応可能「中程度」	学習を用いているため対応できるので「高い」

注：385 ページ分のテストコレクションを処理したときの結果である．

実験の対象として，クエリ「ノートパソコン」(以下，クエリには「」を付ける)を Google で検索したときの検索結果を基に 385 ページ分の HTML ソースからなるテストコレクション(分類の性能を評価するデータセット)を作成した(「ノートパソコン」の検索結果は約 800 ページになり，その中から 6 つのカテゴリーを代表すると考えられた 385 ページ分の HTML ソースを選んだ)．テストコレクションの HTML ソースを予め 6 つのカテゴリー(enterprise, news, shop, diary, bbs，others)に分けて，評価実験を行うことにした．テストコレクションの内訳を表 4-1 に示す．

4.2. 分類器の分類手順

図 4-1 と以下に 3 つの分類器による Web の自動分類の分類手順について解説する．

1. Internet からクローラ[3]により URL 集合を取得する．
2. URL から HTML ソースを取得する．
3. HTML ソースを分類器に入力する．
4. 各 HTML ソースの特徴量を抽出し，分類を行う．
5. 出力として各 URL に対するカテゴリーを得る．

4.3. 評価実験の結果

3 つの手法それぞれについての評価実験の結果を表 4-2 に示す．以下，評価実験の結果から得られた各手法の性質について述べる．

判別分析分類器は，分類精度が高くシステムの負担が少なかったので，大規模なページ数の処理に向いている．しかしスケーラビリティが低いため対象領域を広くすれば，実用的な利用に耐えうる精度は得られない．Rule Base 分類器は，本実験において最も高精度で，ユーザの負担がなく，高速な処理が可能であったため，大規模なページ数の処理に向いていると考えられた．Web ページ間で表現の違いを Rule として全て書き足さなければいけないため，スケーラビリティはあまり高くなかった．しかし，Rule の作成が容易であるため，対象領域と開発者の負担の兼ね合いによって十分に実用的に使用できるものであると考えられる．Naïve Bayes 分類器は，各カテゴリーの分類精度が最

[3]中塚康介氏(京都大学大学院 情報学研究科 社会情報学専攻 博士課程)が作成した Google 専科を使用した

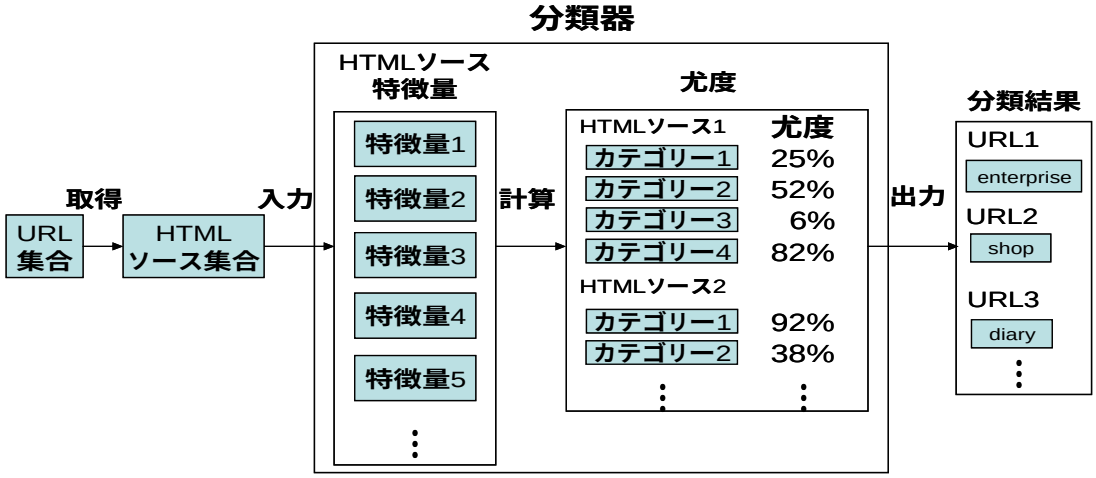

図 4-1　Web の自動分類の分類手順の概念図

も安定していたため，スケーラビリティが高かった．処理する時間が十分にあり，高精度な分類を行いたい場合に有効であると考えられる．

以上より，5 章で解説する応用実験では，Rule Base 分類器(分類精度が高く，ユーザの負担がない，高速処理)を利用する．利用シーンの性質によって，他の 2 つの手法も有効に活用できる場合が存在すると考えられる．

5. Rule Base 分類器による応用実験

本章では，Rule Base 分類器を用いて，Web の自動分類が Web 空間の利活用において, どのように有効に貢献できるのかを検証する．以下，応用実験の対象，手順，結果について述べていく．

5.1. Rule Base 分類器を用いた応用実験の対象

航空会社とノートパソコンに関する Web ページの時系列データをアーカイブしていき，各カテゴリーのページ数の推移を調査することにした．調査期間は 2004 年 12 月 21 日から 2005 年 1 月 31 日まで毎日である．航空会社については「ANA」，ノートパソコンについては「VAIO」の Google での検索結果をアーカイブしていった．ページ数は Google の検索結果で取得可能な最大ページ数を取得した．以下，「ANA」，「VAIO」に関する調査目的を示す．

- 「ANA」:2004 年 12 月 26 日に発生したスマトラ島沖地震という突発的な事件に対して，航空会社 ANA に関する Web ページ数の各カテゴリーの変動を調査する．Web コミュニティが見る ANA に対する情報の構成がスマトラ島沖地震に対してどのように変わったかを見る．
- 「VAIO」：2005 年 1 月 5 日に発表，1 月 15 日に販売された新製品 VAIO に関する Web ページ数の各カテゴリーの変動を調査する．Web コミュニティが見る VAIO に対する情報の構成が新製品の発表・販売によってどのように変わったかを見る．

5.2. Rule Base 分類器を用いた応用実験の手順

「ANA」,「VAIO」の応用実験で行った Rule Base による Web の自動分類の手順を示す．

1. Internet からクローラで URL 集合を取得する．
2. URL 集合から HTML ソースを取得する．
3. HTML ソースに各出来事に関連するキーワード(津波，地震，製品名，製品型番号など)が存在するか調べ, 分類対象となる HTML ソースを選別する．
4. HTML ソースから特徴量を抽出し，URL の各カテゴリーの得点を計算する．
5. 各 URL のカテゴリーを出力する．

5.3.　応用実験の結果

以下に「ANA」,「VAIO」の応用実験の結果を示す．

5.3.1. 航空会社「ANA」の応用実験の結果

図 5-1 に「ANA」に関する 2004 年 12 月 21 日から 2005 年 1 月 31 日の各カテゴリーのページ数の推移を示す．ページ数は 2004 年 1 月 26 日なら前後 2 日の合計 5 日間の平均, すなわち 2004 年 1 月 24 日から 1

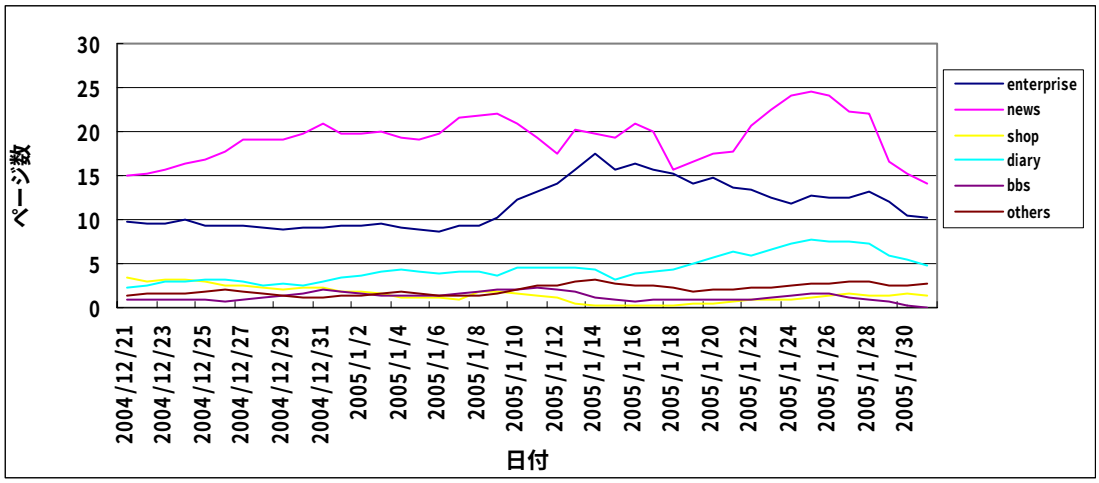

図 5-1 「ANA」の応用実験の結果

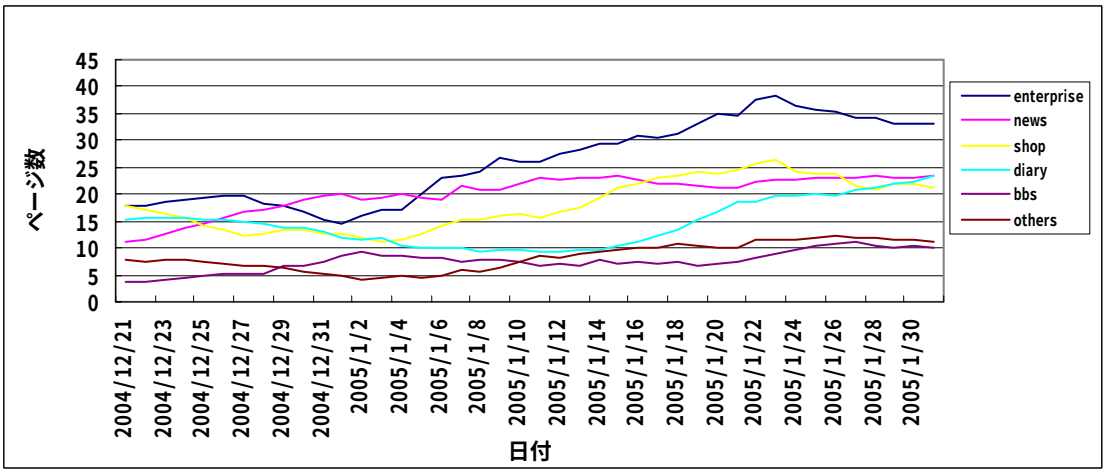

図 5-2 「VAIO」の応用実験の結果

月 28 日までのページ数の平均を表示してある．これは 1 日ごとのページ数の変動が激しかったので，正規化するために行った．

図 5-1 より，スマトラ島沖地震が発生した翌日 2004 年 12 月 27 日からは news に属する Web ページの数が増加傾向にあるのは，事件の状況を解説する記事が増えていたからであった．2005 年 1 月 9 日以降の enterprise の Web ページの増加は,被災地への旅行ツアーが再開されたことを示していた．2005 年 1 月 20 日から diary の Web ページの数が増えだしていたのは，各航空会社の対応や被害の状況に関するコメントの日記が増えていたことを示していた．

5.3.2. ノートパソコン「VAIO」の応用実験の結果

図 5-2 に「VAIO」に関する 2004 年 12 月 21 日から 2005 年 1 月 31 日の各カテゴリーのページ数の推移を示す．「ANA」のときと同様，5 日間平均のページ数を表示している．

図 5-2 より, 2005 年 1 月 5 日の新製品発表のからカテゴリーenterpeise の Web ページの数が増加していることが観測されている．これは新製品を紹介するページが増えていたことを表していた．年末から news の Web ページの数が増えていたのは， SONY が新製品の開発に着手していることを扱った記事が増えていたためである．新製品の一部が 2005 年 1 月 15 日に発売となったので, shop では翌日の 2005 年 1 月 14 日あ

たりから Web ページの数が増加していた．また diary に関して言えば 2005 年 1 月 17 日以降の Web ページの数の増大は，新製品に対するコメントが書かれている日記が増えていたことを表していた．

5.4. 考察

「ANA」の応用実験では，突発的な事件による各カテゴリーの Web ページの数の反応を把握することができた．news と diary は事件に関する Web ページが増えていたのに対し，enterprise は旅行ツアーの再開を表し，各カテゴリーでの反応の意味の違いも観測することができた．「VAIO」の応用実験では，予めわかっている出来事に関する Web ページの反応を把握することができた．news では，この出来事が起こる前にページ数が増え，一般の人々に新製品の開発に関する情報を伝えていた．一方，enterprise は新製品の発表時に，shop,diary は新製品の販売後にそのページ数が増えていた．このように予め起こる出来事がわかっているものに関して言えば，その前後で各々カテゴリーのページがそれぞれの意図で数を増やしていることが観測された．このことから Web の自動分類は，社会的な事件が発生したときに各カテゴリーに属するページ数の変動を把握することができていると考えられる．このような調査を手動で行うことは，膨大な時間的コストがかかるので，Web の自動分類手法は，実用的な使用に対して有効であるということができる．また分析を一歩進めて，相関関数をとると Web 空間上での情報の伝播についての時定数の解析といった高度な Web の利活用につなげることが期待される．また企業の HP のアクセスログから得られる参照元の URL によって，Web ページを自動分類し，Web コミュニティ(enterprise 以外のカテゴリーに属する Web ページ)と企業(enterprise)との良好な関係を保つ戦略を練るときの基礎資料を提供するツールとしての利用も期待される．以上より Web の自動分類は，高度な Web の利活用を行うための分析のツールとして，単独または他の技術と連携することで有効に活用できる可能性があると考えられる．

6. おわりに

本稿では，Web の自動分類によって社会的に大きな出来事が起こったときの Web 空間上での Web ページ数の反応を見ていった．判別分析，Rule Base，Naïve Bayes の 3 つの Web の自動分類器を作成し，それぞれの性質を見た．そして本研究の利用シーンに最も合致した Rule Base 分類器を用いて，スマトラ島沖地震とノートパソコン新製品の発表・販売の出来事に関するカテゴリーごとの Web ページの数の反応を調査した．その結果，Web の自動分類器は時間的に低コストで，Web 空間上の各カテゴリーの Web ページ数の反応の違いを把握するのに役立つことがわかった．

参考文献

[1] Vivisimo：http://vivisimo.com/

[2] Teoma：http://www.teoma.com/

[3] Grokker：http://www.grokker.com/

[4] Lawrence Kai Shih and David R.Larger, “ Using URLs and Layout for Web Classification Tasks ”, Proceedings of the 13th international conference on World Wide Web,2004.

[5] 松田，福島：文書タイプ分類による問題解決向き WWW 検索システムの開発と評価，情報処理学会研究報告，99-FI-53，pp.9-22(1999)．

[6] C. Pavael, C. Marco, M. Edleno, Z. Niveio, R. Berthier, G. A. Marcos：Combining Link-Based and Content-Based Methods for Web Document Classification , Proceedings of the twelfth international conference on Information and Knowledge management(2003).

[7] JP Glossary/Bayesian：
http://popfile.sourceforge.net/cgi-bin/wiki.pl?JP_Glossary/Bayesian