自然言語処理 118－14
（1997. 3. 22）

WWWページの分類におけるテキストの特徴分析手法

落谷 亮
ochi@flab.fujitsu.co.jp
† 富士通研究所
† 〒211-88 川崎市中原区上小田中 4-1-1

本報告では、従来より情報検索の分野で利用されてきた特徴ベクトルと類似度計算の手法に基づきWWWページを対象に行なった分類実験について述べる。

WWWページやオフィスの雑文書等の分類整理では分類用カテゴリ特徴の訓練データの作成が問題になるが、ここでは、数多くのWWWサイトで提供されているインデックス集やリンク集等、人手で行なわれた分類の結果をカテゴリ特徴訓練データとして用い、単語、形態素、bigram等の特徴素別の分類結果や、リンク情報の利用による特徴収集等について報告する。

Text Property Analysis for classifying WWW Pages

Ryo OCHITANI

† Fujitsu Laboratories Ltd.
† 4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, Japan 211-88

The result of our experiments on the automatic text classification of WWW pages using the IR techniques, such as property vectors and similarity measures, is described in this report. One of the problems to classifying general texts, such as WWW pages and office documents, is to get the trainging data to calculate the categories property. Our experiments used human coded indexes offered by many Web sites. Several results based on the different index terms, words, morphemes and bigrams of morphemes and the use of page links to collect property information are shown.

1 はじめに

イントラネットの普及により企業内の文書情報もネットワーク上に載せることが容易になり、それにつれて企業内の特定のグループの人間の間でしかアクセス出来なかったような文書等がサーバの上に置いてグループ外に公開されたり、オフィス内の雑文書など様々の文書が検索の結果多量に得られるなど、内容の定まらないテキストデータの山を前に必要な情報を取り出すというような機会も増加している。

また、ワープロなどの高性能・大容量化の影響で、これまでに積もり積もった未整理の文書など分類整理の必要なテキストを沢山抱えるユーザが増えている。これらの文書は自分の私的な文書であっても本人にも内容が思い出せない等、他から入手した他人の文書同様の扱いを必要とすることも多い。

この様な状況を反映して、一般の多様な文書を対象にした文書の分類、インデックス、要約作成などの重要性の認識が現実の問題として企業内や個人レベルまで急速に高まってきている。

例えば、インターネットの世界では、Yahoo 等を代表とする組織的で大規模なインデックス集や個人規模でのリンク集など様々な分類情報やコメント情報等を手作業で作る努力が多くの場所で見掛けられ、作成された分類や索引情報は情報の溢れるWWW の中で必要な情報を見つける際に有力な手段であることを実感することは多い。

しかし、これらのテキスト情報の分類、索引付けやアノテーションなど整理の作業には多大の労力が掛かり、イントラネットに置かれた社内文書や個人の文書など、インターネット程の必然性が明確でなく、また雑文書の様な分類や索引作成の労力が見合わないとも考えられるれるような文書については、なかなか情報が整えられないのが現状である。

我々の研究で目標としている文書情報管理システム（図1）では、このような利用価値は高いが作成コストの高い分類、索引、アノテーションなどの文書の周辺情報や自動分類、自動抜粋ツールなどの文書処理ツール類を文書リポジトリの統合的な枠組の中に取り込み一括して管理することで、文書及び文書の周辺情報も含めて再利用が可能で情報の共有が容易な文書処理環境を実現しようと考えている。

例えば、分類用の特徴情報を共有し、ある文書グループを対象に対して作られた分類特徴を、似たような別のテキストの分類の際に活用できれば、新たな対象テキストに対しての特徴情報の収集作業を軽減することができると考える。

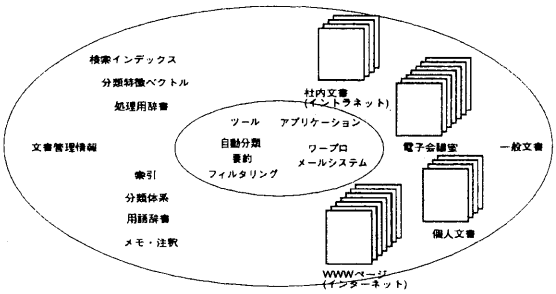

図 1: 文書リポジトリ

このような観点からすると、WWW に流通するガイダンス情報やインデックス情報等、手作業の分類結果や要約結果の貴重な情報を、分類、要約、その他の自動処理システムのシステムに与える訓練セットとして利用することが出来れば、イントラネットや個人の収集した情報の整理にも役立つのではないかと考えられる。

しかしながら、WWW の文書は直観的にみて非常にばらつきの多いテキストであると思われ、実際どの程度の利用価値があるのかは不明であった。

そこで、実際にページデータを集め、情報検索の分野で従来より用いられている特徴ベクトルと類似度の計算の手法により分類処理を行なうという方法[3]で、WWW ページのような、まとまりのない大小のテキストが入り混じったデータに対しての自動分類処理が、どの程度実現可能かを調べてみてみようというのが、ここで報告する実験の目標である。

実験の中で用いた手法としては、テキストからの特徴素抽出として、単語、形態素、形態素の bigram を用いて分類実験をそれぞれ行ない、また、ページのリンクを利用した特徴情報の収集などについても実験を行なった。

2 対象データ

本実験で分類カテゴリの特徴訓練データとしているのは、WWW に見られるインデックス集やリンク集などの人間が分類したデータであり、WWW 上には検索サービスと組合わせられた大規模なナビゲーションのガイドシステムから、個人の集めたリンク集まで様々なレベルの情報が存在し利用できるが、今回の実験では訓練データと評価データを多量に集める必要があるため、情報がまとまって入手可能である Yahoo を初めとする大規模なインデックスサービスの情報を利用した。

この様なサービスでの分類では、ユーザによるナビゲーションを助けるための体系的な構造として、最上位の大分類カテゴリをスタート地点として階層構造の下位のカテゴリへリンクを辿ることによって詳細な情報へ移ることが出来るように緩い包含関係のある体系が用意されている。

このような分類体系を利用した分類実験として、最上位の十数個の緩い大分類カテゴリをまとまりとする大まかな分類処理と、数千に分かれる下位の詳細カテゴリをまとまりとする詳細分類処理の2つのケースに分けて実験を別に行なった。最上位の十数個のカテゴリに対する実験では、上位のカテゴリは下位のカテゴリの特徴を単純に包含すると仮定し下位カテゴリからリンクされていページの特徴ベクトルを全て最上位のカテゴリに加えることで最上位カテゴリの特徴を作成した。

カテゴリの階層構造はWWWがハイパーテキストである点を活かして、階層は木構造ではなくネットワーク状に構成されている。従って、複数の上位カテゴリからリンクされた下位カテゴリも存在し、分類の際には、これらのカテゴリ特徴は両方の上位カテゴリが包含すると仮定した。

実験にあたっては、分類体系のカテゴリページとそれぞれのカテゴリからリンクされたページを合わせて、約4万ページのページデータを収集し、このページデータを訓練セットとテスト用セットに分け特徴ベクトルの作成、分類実験を行なった。

実際に収集したインデックス集のカテゴリデータをみると収集されている情報に非常に偏りが大きいのが分かる。表1に大分類カテゴリ毎のページ数の比率を示す。

表 1: 大分類と対応ページ

大分類カテゴリ	比率 %
ビジネス	32.4
エンターテインメント	23.9
地域情報	18.1
スポーツ、レジャー	6.0
コンピュータ	5.7
アート	3.0
教育	2.8
科学	2.4
社会、文化	2.3
健康	1.6
政治	0.56

大分類カテゴリでは「ビジネス」関連のページへのリンクが1万件以上あるのに対し、「健康」や「政治」といった分類へのリンクは数百といった所でありカテゴリ毎に含まれる特徴情報の量に差は大きい。

詳細分類についてみると(表2)、「エンターテインメント・個人ページ」へのインデックスの様に一つの分類に5千のURLがリンクされているページもあるが、「地域情報・長崎・イベント」例の様にカテゴリに1つのリンクしか持たないカテゴリも多い。

表 2: 詳細分類と対応ページ

詳細分類	比率 %
エンターテインメント・個人	13.6
エンターテインメント・音楽・アーティスト	0.8
ビジネス・コンピュータ・ソフトウェア	0.8
ビジネス・インターネット・プロバイダ	0.7
・・・	・・・
アート・イベント・フェスティバル	0.002

次に表3にカテゴリにより分類されているページの特徴素数を示す。特徴素数の平均から見るかぎりはページ当たりの情報は少ない。

表 3: 文書当りの平均特徴数

特徴素	平均数
形態素	79 形態素
連語処理	58 語
bigram	35 ペア

1ページ当たりの特徴素数は少ないが、特徴素がスパースな性質は高く、ページ数の増加につれて図2に示す様に増加する。このグラフからは増加からみると形態素は特徴素数の増加の面では有利であり、bigramがその次に続くと考えられる。

3 実験方法

実験手法として特徴的な部分だけ以下に挙げる。

1. ページ収集および前処理

   インデックスサービスやリンク集のサイトから索引情報のページを取り込み、(分類カテゴリ, リンク先ページのURL) の組を取り出す。このURLを元に、実際のリンク先ページを取

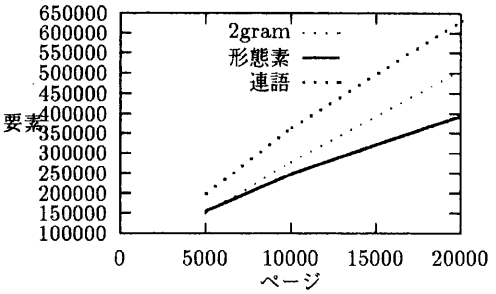

図 2: 各特徴素数の増加

りだし、GIF などのイメージや Postscript 等テキスト以外のデータは取り除き、リンク情報とテキストを分離する。

2. 特徴ベクトル及び分類カテゴリの計算

分離したテキスト部分を形態素解析に掛け、形態素、形態素の bigram（自立語、接頭語、接尾語の連続からペアを作る）、連語の集計を行ない特徴ベクトルを作成する。

各ページの特徴ベクトルは先の特徴素をタームとした $tf \cdot idf$ によって与える。

(単語1の$tfidf$, 単語2の$tfidf$, ..., 単語nの$tfidf$)

3. カテゴリ特徴訓練処理

インデックスサービスのページから抽出した URL と分類カテゴリのペアを訓練セットとして用いるが、大分類カテゴリを用いた実験では、各大分類カテゴリの特徴ベクトルは、そのカテゴリの下位のカテゴリからリンクされている全てのページの特徴を全て加えたものとする。

大分類カテゴリの特徴ベクトル =

$\sum$ カテゴリと下位カテゴリの文書特徴ベクトル

詳細分類を用いた実験では、各詳細分類カテゴリのページに直接リンクされたページの特徴ベクトルを加え詳細分類カテゴリの特徴ベクトルとする。

詳細分類カテゴリの特徴ベクトル =

$\sum$ カテゴリに属する文書特徴ベクトル

4. 特徴ベクトルとカテゴリ特徴間の類似度は *cosine* 類似度を用いて計算する。

カテゴリ C_i 及び文書 D_j に含まれる特徴素 T の $tfidf$ をそれぞれ T_{it}、T_{jt} とすると類似度は以下の式で計算される。

$$cosine(C_i, D_j) = \frac{\sum_{t=1}^{n}(T_{it} \times T_{jt})}{\sqrt{\sum_{t=1}^{n} T_{it}^2 \sum_{t=1}^{n} T_{jt}^2}}$$

5. リンク先のページ収集

あるページからリンクされているリンク先のページは、もとのページに近い内容を表していると推測できる。このような観点からリンク先のページ特徴を、カテゴリ特徴の作成やカテゴリ類似度判定の際に元のページの特徴に追加するという実験である。

リンク先情報も含めた特徴ベクトルの作成を行なう場合には、リンク情報からリンク先のページを取り込み、前処理、特徴ベクトルの計算を行ない最初のページの特徴ベクトルとの平均を取る。

6. 実験結果の評価

カテゴリ情報と対で収集したページのうち、訓練セットに選んだページとは別のページを評価セットとして選び、以下の再現率、適合率により分類結果の評価を行なう。

- 再現率

$$\frac{\text{正しく割り付けたカテゴリ数}}{\text{評価セットの正解カテゴリ数}}$$

- 適合率

$$\frac{\text{正しく割り付けたカテゴリ数}}{\text{割り付けたカテゴリ数}}$$

4 実験結果

4.1 特徴素別の分類結果

この実験では、訓練セットとして 1 万ページ評価セットとして約 4200 ページ分の特徴ベクトルをランダムに選び訓練・評価のデータとした。

1 万ページを 14 個の大分類カテゴリの訓練セットとして用い、評価セットのページをその大分類に分類し正解カテゴリとの照合により評価を行なった。

図 3 に特徴素の取り方を形態素、形態素の bigram、連語のそれぞれに変えて行なった実験の結果を示す。

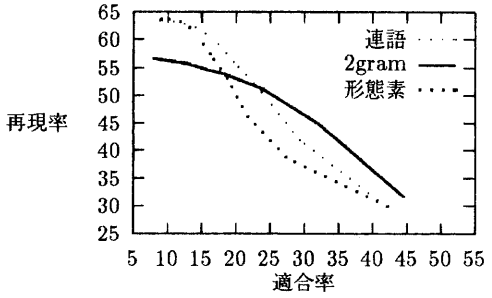

図 3: 特徴素別の適合率 / 再現率

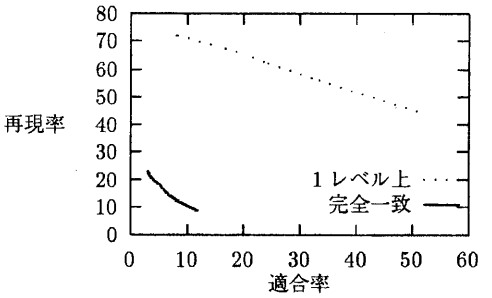

図 4: 詳細カテゴリによる実験

結果からは形態素の bigram によるものが、若干ではあるが再現率、適合率共に良い値を示すことが多い。これは西野 [2] の特許文に対する特徴抽出の結果と同様の傾向を示している。

実際の特徴ベクトルや分類の際の類似度計算の状況をみると、対象テキスト中の空白や記号などにより形態素解析が失敗し、細切れの情報を多く返すため、bigram の計算の際にこれらのノイズが自然に除去されている可能性を示していると考えられる。

4.2 詳細カテゴリへの分類結果

この実験では詳細カテゴリへの分類を行なった。訓練セットとして先の 1 万ページの訓練セットの中から「地域情報」、「エンターテインメント」のカテゴリのサブカテゴリ約 1100 詳細カテゴリを選び、それらの詳細カテゴリに属するもの 7800 ページを訓練セットとして詳細カテゴリの特徴を作成した。

評価セットとしては、先の約 4200 ページ分の特徴素から、やはり、「地域情報」、「エンターテインメント」にのサブカテゴリに属するものを選びデータとした。

結果を図 4 に示す。詳細カテゴリが完全一致するものは訓練セットが小さ過ぎるためか非常に少ない。そこで、評価の際のカテゴリの一致判定を一レベル上のクラスまで合えば正解に含めたものを計算してみた。これは、例えば「合衆国・合衆国の州・ニューヨーク」のカテゴリ判定の際に「合衆国・合衆国の州」まで合えば正解とする評価基準である。

4.3 リンク探索による分類結果

図 5 に、連語を特徴素とした場合について、テキスト特徴とカテゴリ特徴の間の類似度計算の際にリンク先のページの特徴を追加した実験の結果を示す。図 6 は、カテゴリ特徴の作成の際に同様にリンク先特徴を追加した実験の結果である。

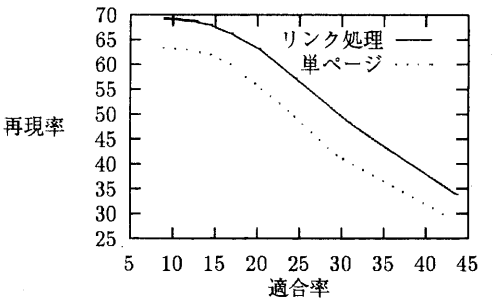

図 5: リンクによるテキスト特徴の補間

両方の場合につき、再現率、適合率ともに若干の向上が見られた。

5 検討

実験結果として示したように、大分類カテゴリへの分類結果を見る限りは、再現率、適合率ともに高いとは言えない。しかし、詳細分類カテゴリの分類結果で示されるように、分類体系の上で 1 つ上のカテゴリを推定するような場合には若干は結果が良

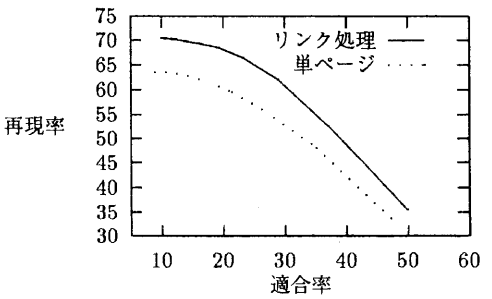

図 6: リンクによるカテゴリ特徴の補間

い。

また、1ページの情報だけでなくページからリンクされている先のページも辿ることでも結果は少し良くなる。

今回の実験ではカテゴリ割り付けそのものは単純にk個のカテゴリを点の高いものから順に割り付けているが、西野[1]の信頼度に基づくカテゴリ割り付けを用いれば、今回のデータの様なカテゴリ間の特徴情報量に偏りが大きい場合の結果は改善出来ると考えられる。

6 結び

一般に多く見られる雑文書の典型としてのWWWページをIR手法に基づいた処理により分類する実験結果について述べた。

最初に示した様に、これまでの実験で得たような特徴情報を、社内文書など分類特徴の訓練データの少ない分野の文書に利用することも考えており、量は少ないが社内の検索サーバから入手した数百ページのデータについて分類実験を行なってみている。

結果を大雑把に見た範囲では、約半分程度ののデータが分類カテゴリと関連あるデータを示しており、例えば「新製品のニュースリリース」が「ビジネス」「新着情報」がカテゴリ「ニュース」に割り付けられるというような傾向が見られた。分類精度は高くないが、検索を大雑把に仕分けるなどのには利用出来る可能性はあると考える。

今回の分類実験は分類可能性の検討のために高速のサーバマシン上で行なっている。しかし現実的には、このような一般文書の分類はさまざまな局面で必要であると考えられ、特に個人で入手した文書を大雑把に分類したいなどのニーズも予想されるので、文書特徴データの圧縮や高速の特徴検索など、より規模の小さなパソコン等で実行可能な分類処理の必要性を感じている。

また、個人文書のような小規模のデータに対する分類処理では、データだけを自動クラスタリングアルゴリズムに与えて分類することもデータの規模としては可能であると考えられる。実際に小規模なクラスタリング実験の結果からは、クラスタリング結果を活用するには、我々が通常理解しているような分類に当てはめる操作が必要と考えられ、本実験で抽出したような人間の分類特徴と自動クラスタリングの結果の照合処理などの実験についても検討中である。

7 謝辞

日頃よりインターネット上での検索やナビゲーションの助けとして有効に活用させて頂き、また、本実験の材料としてシステムの訓練、評価の元データとさせていただいたインデックス集やリンク集を作成されている多くの方々、Yahoo、その他の方々の努力に感謝する。

8 参考文献

1. 西野文人：テキスト分類のためのカテゴリ割り付け戦略、情処研報、NL106-3,pp.13-18(1995)
2. 西野文人：日本語テキスト分類における特徴素抽出、情処研報、NL112-14,pp.95-102(1996)
3. Salton, Gerard and McGiLL,Micael J.：Introduction to Modern Information Retrieval, MacGraw-Hill,1983.