ウェブページの種別を識別する論理式の提案とその市場調査への応用*

1T－6

菊池 崇宏　菊池 浩明　中西 祥八郎†

東海大学工学部‡

1 はじめに

インターネットの普及に伴い、個人の趣味のページから企業概要まで、種々のWebページが公開されるようになってきた。それらの膨大なデータの中から、意図するページを抽出する検査エンジンの技術も成熟してきた。そこで、本研究では、Webページの統計情報を測定することで、インターネットの市場調査への応用を試みる。例えば、個人のページの内、「サッカー」を含むページの数を調べれば、「サッカー」の人気の度合が直接測定出来る。ところが、個人のホームページも一様でないので、単一のキーワードだけでは特定しきれない。そこで、本稿では、論理決定木による学習を導入し、意図する種類のページだけを抽出するキーワードの論理式を同定する。得られた式を用いて実際に市場調査を行い、文献によるデータと比較検討する。なお、本稿における実験データは1998年12月に測定したものである。

2 提案調査法

2.1 検索エンジンの予備調査

適切な検索エンジンを探す為、メジャーな20種のサービスを調査した。その結果、最も登録件数が多く、論理型検索を提供しているキーワード型検索エンジンGoo[2]を利用することに決めた。

2.2 学習データの定義

Webページの種類を識別する特徴的なキーワードを探る為、あるキーワードを含むページをサンプリングする。サンプリングされた全ページについて、種類分けを行ない、いくつかの特徴的なキーワードについての包含関係を調べ、それを学習データとする。

キーワード「崇宏」でサンプリングした時のページの種類を表2に示す。7種類への識別は、著者らの主観で判断した。この内、個人のページに識別されたページとキーワードとの関係の一部を表1に示す。

2.3 論理決定木の同定

学習データにID3[3]などのアルゴリズムを適用し、識別種類の各々について論理決定木Fを定める。ただし、学習データによっては、含まれたキーワードが全く一致するのに、異なる種類に識別される矛盾した部分が生じることもある。

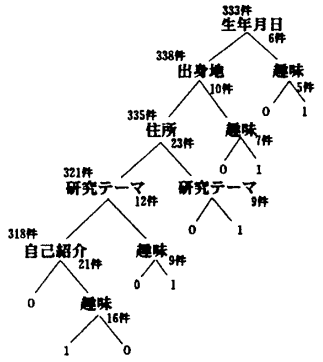

図1: 個人のページの論理決定木

図1は、表1の学習データについて定めた論理決定木である。ノードのキーワードを含むならば右、含まないならば左の枝に遷移する。図の上に、その部分木に属する件数を表示している。

この論理決定木を論理式に展開し、検索エンジンにかけると、意図する「個人のページ」よりも多くのページがヒットする。会合などの参加者リストなどが含まれるからである。そこで、それらを排除するキーワード「$\overline{\text{一覧表}}$」と「$\overline{\text{参加者}}$」を論理積し、最終的に論理式

$$
\begin{aligned}
F = {} & (\text{生年月日} \wedge \text{趣味} \vee \overline{\text{生年月日}} \wedge (\text{出身地} \wedge \text{趣味} \\
& \vee \overline{\text{出身地}} \wedge (\text{住所} \wedge \text{研究テーマ} \vee \overline{\text{住所}} \wedge \\
& \text{研究テーマ} \wedge \text{趣味} \vee \overline{\text{研究テーマ}} \\
& \wedge \text{自己紹介} \wedge \overline{\text{趣味}}))) \wedge \overline{\text{一覧表}} \wedge \overline{\text{参加者}}
\end{aligned}
$$

*Logical Approach for Identification of Web pages and Application to Marketing

†Takahiro Kikuchi, Hiroaki Kikuchi, and Shohachiro Nakanishi

‡Tokai University

表 1: 個人のホームページの例 (学習データ)

No.	URL	趣味	血液	生年月日	住所	出身地	研究テーマ	自己紹介
1	http://www.tokai.ac.jp/t98252tk/index-j.html	1	1	1	1	1		
2	http://www.tokai.ac.jp/Mem/taka/index.html				1		1	
3	http://www.indes.co.jp/prv/kawai.html	1	1	1		1		
4	http://www.info.or.jp/ōkonogi/profile.html	1	1	1	1	1		1
5	http://www.edu.cu.ac.jp/d̃23009/index.html	1	1			1		1
6	http://www.japan.ac.jp/ mura/jiko.html	1	1	1	1	1		1
7	http://www.kawalab.ac.jp/ yanagi	1					1	

を定める。ただし，Gooでは否定とかっこの組合せがサポートされていない為、最小項展開の式を用いる必要がある。この式でサンプリングデータ (「崇宏」) における16件の個人ホームページ中14件のページが識別された.

2.4 他の学習データによる検証

同定された式Fの正当性を確かめる為、他のサンプリングデータにもFを適用して正当率を求める。この検証結果を表3に示す。サンプリングキーワードには、学習データと同じ規模のサンプリング数となるもの5つを選んだ。Fは約70%の確からしさで、「個人ページ」を識別していた。

2.5 市場調査

サンプリングから得られた式Fを、検索エンジンの全登録ページに適用する。調査したい項目 $x_1,\ldots,x_m$ について、$F\wedge x_1,\ldots,F\wedge x_m$のヒット件数を求めて調査結果とする。

GooにおけるFで識別される個人のホームページは125,666件ある．その内、各種スポーツ ($x_1,\ldots,x_m$) のヒット件数を表4に示す。単一のページが複数のスポーツ名を含んでいる場合があるので、合計は12万を越している。

表 2: サンプリングデータにおけるページの種類

ホームページの種類	件数
個人のページ	23件
研究関連のページ	116件
スポーツのページ	70件
会社関連のページ	16件
リンクリストのページ	51件
その他	40件
計	317件

表 3: 他のサンプリングデータにおける正解率

サンプリングキーワード	正解/総件数	割合
耕司	42/50	84%
隆一郎	11/21	52%
辰則	10/14	71%
剛志	59/85	69%
啓太	18/24	75%
平均正解率		70.2%

表 4: インターネットにおけるヒット件数

種目	件数	文献による順位
スキー	6,275件	8
野球	5,931件	1
サッカー	5,307件	7
テニス	5,237件	5
ゴルフ	3,148件	5
水泳	2,539件	11
バスケットボール	1,413件	9
バレーボール	1,195件	9
マラソン	764件	3
登山	741件	12
相撲	536件	4

3 考察

表??では、比較の為に、文献 [1] における「日本人に人気のあるスポーツ」の順位も示している (調査は、3000人のランダムサンプリングにより、2項目までの選択が許されていた)。文献では8位のスキーがインターネットでは1位である様に、両者は必ずしも一致していない。この原因として、次の点があげられる。

1. インターネットを利用する年齢層の偏り
2. 調査法の不統一 (選択出来る項目数などの違い)
3. 論理式Fの同定誤差

4 おわりに

インターネットのWebページを利用した市場調査方法を提案した。実験により、意味のある統計量を測定出来ることが示されたが、従来の統計手法による結果には一致しなかった。しかしながら、低コストで機械的にデータ集計行なう利点を持つインターネットには、潜在的な市場調査への可能性を感じる.

参考文献

[1] 平成9年版 世論調査年監-全国世論調査の現況-, (総理府) 内閣総理大臣官房広報室編,pp.480, 1998

[2] 検索エンジン Goo、(http://www.goo.ne.jp)

[3] Quinlan,J.R., Induction of decision trees,Machine Learning, 1(1),pp.81-106,1986